VYSOKÁ ŠKOLA BÁŇSKÁ – TECHNICKÁ UNIVERZITA OSTRAVA
EKONOMICKÁ FAKULTA

KATEDRA FINANCÍ

Modelování četnosti pojistných škod v havarijním pojištění

Modelling of Claim Frequency for a Motor Hull Insurance Portfolio

Student: Bc. Ester Lysková

Vedoucí diplomové práce: Ing. Jiří Valecký, Ph.D.

Ostrava 2016

VŠB - Technická univerzita Ostrava
Ekonomická fakulta
Katedra financí

Zadání diplomové práce

Student: **Bc. Ester Lysková**

Studijní program: N6202 Hospodářská politika a správa

Studijní obor: 6202T010 Finance

Téma: Modelování četnosti pojistných škod v havarijním pojištění
Modelling of Claim Frequency for a Motor Hull Insurance Portfolio

Jazyk vypracování: čeština

Zásady pro vypracování:

1. Úvod
2. Charakteristika rizika v pojišťovnictví
3. Charakteristika modelů četnosti
4. Odhad a vyhodnocení regresního modelu
5. Závěr

Seznam použité literatury
Seznam zkratek
Prohlášení o využití výsledků diplomové práce
Seznam příloh
Přílohy

Seznam doporučené odborné literatury:

HARDIN, James W. and Joseph M. HILBE. *Generalized Linear Models and Extensions.* 2nd ed. College Station: Stata Press, 2007. 387 s. ISBN 978-1-59718-014-6.
JONG, Piet de and Gillian Z. HELLER. *Generalized Linear Models for Insurance Data.* Cambridge: Cambridge University Press, 2008. 196 s. ISBN 978-0-521-87914-9.
OHLSSON, Esbjörn and Björn JOHANSSON. *Non-Life Insurance Pricing with Generalized Linear Models.* Berlin: Springer, 2010. 174 s. ISBN 978-3-642-10790-0.

Formální náležitosti a rozsah diplomové práce stanoví pokyny pro vypracování zveřejněné na webových stránkách fakulty.

Vedoucí diplomové práce: **Ing. Jiří Valecký, Ph.D.**

Datum zadání: 20.11.2015
Datum odevzdání: 22.04.2016

Ing. Iveta Ratmanová, Ph.D.
vedoucí katedry

prof. Dr. Ing. Dana Dluhošová
děkanka fakulty

Prohlašuji, že jsem celou práci, včetně všech příloh, vypracovala samostatně.

V Ostravě dne 22. dubna 2016

Eliška Lysíková

Tímto bych chtěla poděkovat vedoucímu diplomové práce Ing. Jiřímu Valeckému, Ph.D. za vstřícný přístup, dohled, ochotu a věnovaný čas při tvorbě diplomové práce.

Tato diplomová práce vznikla za podpory projektu SGS SP 2015/75 Aplikace zobecněných lineárních modelů v pojišťovnictví a ve financích.

Obsah

1 Úvod

V dnešních časech, kdy na silnicích je mnoho motorových vozidel, často dochází k dopravním nehodám. V zájmu různých ekonomických subjektů je minimalizovat riziko dopravní nehody a s tím související vzniklé škody nebo alespoň přenést riziko na jiný subjekt. K tomu je používáno havarijní pojištění motorových vozidel. Pojistitelé ovšem na pojištění nechtějí prodělat a proto jsou tvořeny modely četnosti pojistných událostí a ty jsou následně používány k určení výše pojistného. Pro odhad rizika u klienta jsou novými technologiemi v pojišťovnictví využívány údaje obsažené v pojistné smlouvě. Vhodným nástrojem pro určení rizikovosti pojištěného jsou regresní modely. Těmito modely je možná identifikace vlivu jednotlivých ukazatelů na četnost pojistných událostí, která je vyjádřena odhadem střední hodnoty počtu pojistných událostí za určité období.

Cílem diplomové práce je modelování četnosti pojistných škod v havarijním pojištění. Pro odhad vhodného modelu je použita *metoda maximální věrohodnosti*.

Diplomová práce je rozdělena do pěti kapitol včetně úvodu a závěru. V první kapitole je představen cíl a obsah práce.

V druhé kapitole jsou popsána rizika v pojišťovnictví a možnosti řízení rizik.

Třetí kapitola je zaměřena na popis metodiky odhadování parametrů modelů, které jsou následně aplikovány na reálných datech v praktické části seminární práce. Nejdříve je vyjasněno, co to je zobecněný lineární model, blíže je popsáno negativně binomické rozdělení, dále je uvedena metoda odhadu parametrů modelu a jeho verifikace.

Na začátku čtvrté kapitoly jsou stručně popsány základní informace o vstupních datech a je určena významnost koeficientů vysvětlujících proměnných pomocí jednofaktorové analýzy. Dále jsou odhadovány koeficienty veličin modelů metodou maximální věrohodnosti a grafické zobrazení predikovaných reziduí odhadovaných modelů. V další podkapitole jsou spojité veličiny kategorizovány a je odhadován model s kategorickými proměnnými. Na závěr praktické části je zhodnocen vliv kategorizovaných veličin na četnost pojistných škod v havarijním pojištění.

V závěru jsou shrnuty všechny předchozí postupy diplomové práce a je napsáno zhodnocení odhadovaných modelů.

2 Charakteristika rizika v pojišťovnictví

Druhá kapitola je zaměřena na obecný popis rizika a s tím souvisejícího pojištění. V celé druhé části práce je čerpáno z odborné literatury Cipra (2006a) a Cipra (2006b). Kapitola je rozdělena na několik podkapitol. První podkapitole je věnována obecným teoretickým poznatkům jako charakterizaci, klasifikaci rizika a základním pojmům v pojištění. Další podkapitola je zaměřena na stručný popis havarijního pojištění a v poslední podkapitole je zmíněno, co to je risk management.

2.1 Charakteristika rizika a pojištění

Ve světě, ve kterém žijeme, na nás číhá plno nejistot a nahodilých událostí. Ekonomický subjekt je stále vystavován nějakému nebezpečí vzniku škody. Nebezpečím je ovlivňována jistota subjektu a synonymem nebezpečí škody je tedy nejistota. Riziko můžeme chápat jako vznik události s rozdílným výsledkem a určitou pravděpodobností. Důsledky vzniku nepředvídatelných událostí mohou být záporné, ale i kladné. Riziko je měřitelná část nejistoty a měříme ji pomocí pravděpodobnosti. Nesmíme ovšem opomenout existenci tzv. pravé nejistoty, kterou nelze změřit, jelikož u ní nejde určit rozložení pravděpodobnosti. V riziku je zahrnována jak kvalitativní, tak i kvantitativní stránka. Měříme nejen pravděpodobnost vzniku nahodilosti, ale rovněž i rozsah a závažnost možných následků. Je důležité riziko předvídat, posuzovat a samozřejmě i omezovat.

Pojištění je nástrojem eliminace negativních důsledků nahodilosti. Jednou z charakteristik je, že pojištění se zabývá jevy náhodného charakteru a důsledkem nahodilosti je vznik nějaké škody. Existence takového finančního nástroje je motivována eliminací negativních důsledků nahodilých jevů. Náhodné jevy jsou označovány jako pojistná rizika. Dojde-li k jejich realizaci, jedná se o pojistné události, při níž pojišťovna vyplácí pojistné plnění dle ujednaných smluvních podmínek.

Pojistné plnění je řešeno formou finančních náhrad. Jiný typ krytí není pojišťovnou poskytován.

Pojištění můžeme brát jako ochranu proti rizikům. Pojištěný přenese rizika, kdy důsledky z hlediska pojištěného jsou neúnosné, na pojistitele (např. pojišťovnu, penzijní fond, apod.). Soubor všech pojistných smluv je označován jako pojistný kmen. V zájmu každé pojišťovny je mít co největší pojistný kmen.

Pojistná rizika jsou klasifikována dle různých hledisek. V práci se zmíníme o vybraných druzích rizika.

Čisté riziko je náhodného charakteru na rozdíl od uměle vytvořeného *spekulativního rizika* např. riziko spojené se sázením, spekulacemi na burze apod. V čistém riziku je zahrnut pouze *negativní* důsledek nahodilé události, čili záporná odchylka od očekávaného stavu. Předmětem pojištění je pouze čisté riziko. Ekonomické subjekty se tak mohou bránit negativním důsledkům plynoucím z čistého rizika. Čisté riziko dále můžeme rozdělit na subjektivní a objektivní. *Subjektivní* typ čistého rizika je spojen s jednáním lidí. Podstupují-li subjekty morální riziko či jsou neopatrní, jedná se o typ subjektivního rizika. *Objektivní riziko* je dáno objektivními skutečnostmi: např. blesk, přírodní katastrofy nebo věk, profese, pohlaví aj. Nahodilost pojistného rizika je buď *relativní* jako úmrtí, jež určitě nastane, ale neví se kdy. Nahodilost může být rovněž *absolutní,* např. požár.

Morálním rizikem nazýváme situaci, kdy pojištěný preferuje zábranné riziko místo vzniku škody, ovšem se nesmí jednat o pojistný podvod. Morální hazard je rovněž brán jako tendence změny chování pojištěných osob k získání co nejvyššího pojistného plnění z uzavřené pojistné smlouvy.

Do *osobního* rizika počítáme riziko předčasné smrti, tělesného poškození nebo sociální nedostatečnosti.

Dále můžeme rizika rozdělit dle věcného členění na živelní, dopravní, strojní, zemědělská rizika a jiné. **Živelním rizikem** je riziko přímých škod na majetku důsledkem živelních událostí. Mezi vodovodní rizika patří riziko způsobené vodou. Důležitým typem rizik, se kterými budeme pracovat, je **dopravní riziko** např. riziko škod vzniklých dopravním prostředkem nebo v souvislosti s přepravovaným zbožím. Názorným typem pojištění těchto rizik je na příklad havarijní pojištění, kasko letadel nebo kargo pojištění. Další rizika související s majetkem pojištěného je riziko odcizení a vandalství nebo šomážní riziko. **Šomážní riziko** je přerušení provozu nebo výroby v důsledku havárie, výpadku energie atd. **Strojním rizikem** je myšleno riziko havárie nebo poruchy strojního zařízení, které bylo následkem neodborného zacházení, chybné technologie či vady materiálu. Mezi **zemědělská rizika** řadíme riziko ztráty v živočišné nebo rostlinné výrobě v důsledku živelních událostí, úrazu zvířat nebo jarních mrazů. **Odpovědnostním** rizikem je myšleno riziko škod způsobených v důsledku jednání pojištěného na životě či zdraví jiné osoby nebo na cizím majetku. Mezi tato pojištění patří pojištění odpovědnosti za škodu způsobenou provozem

motorového vozidla, výkonem povolání atd. V sociálně- politickém riziku je zahrnováno **válečné riziko**, stávky, embarga atd. Změny ekonomických podmínek, dodavatelsko-odběratelských vztahů patří mezi **obchodně- finanční rizika**. **Úvěrové riziko** je riziko nesplacení dluhů a **devizové riziko** je rizikem s devizovým dopadem. Posledním typem rizik jsou **rizika moderní** jako např. atomové riziko, riziko AIDS, ekologické riziko atd.

Rizika, které pojistitel přebírá od klientů, jsou transformována na pojistně- technické rizika pojistitele. Rizikem je rovněž, že nedojde k vyrovnání mezi přijatým pojistným a vyplaceným pojistným plněním. Dané riziko je měřeno výši variability- *směrodatnou odchylkou* mezi očekávaným stavem a skutečným jevem, jež se projeví ve vyplaceném pojistném plnění.

Z hlediska vzniku náhodných událostí rozlišujeme rozměry rizika na,

- okamžik realizace rizika, když vznik nahodilosti je spojen s nějakým časovým okamžikem;
- výskyt realizace rizika, tento rozměr je spojen pouze s absolutní nahodilým jevem;
- realizace rizika, daný rozměr je sledován pouze u těch rizik, která mohou být realizovatelná jak plně, tak i částečně. Realizace rizika je nastání událostí, jež ohrožuje ekonomický subjekt a vede ke vzniku škody, které mohou mít povahu materiální nebo nemateriální.

Velikost rizika je ovlivňována *četností* a *velikostí škody*. Existují různé vztahy mezi těmito veličinami. Je-li nízká závažnost a nízká četnost, riziko je realizováno zřídka a při realizaci vznikají malé škody. Při vysoké četnosti, ale nízké závažnosti dochází k časté realizaci, ale s malými důsledky. Naopak nízká četnost a vysoká závažnost, představuje velkou hrozbu i pro pojišťovny. K dané události dochází velmi zřídka, ale škody jsou pak obrovské. Ekonomické subjekty se snaží riziku předejít, ale nastane-li nahodilá událost, je pak nutno vzniklé škody krýt. Existuje více možností krytí rizik a to krytí prostřednictvím státu, krytí vlastními zdroji a pojištění. Finanční krytí prostřednictvím státu je možné pouze v případě škod velkého a hromadného rozsahu v důsledku živelních událostí. Dalším typem krytí rizik, je individuální zabezpečení, čili krytí z vlastních zdrojů. Posledním typem krytí rizik je přenesení rizika na pojišťovnu za určitý poplatek.

Jako nejdůležitější charakteristiky rizika považujeme *míru pravděpodobnosti rizika*, že ono nastane, dále *úroveň rizika* a jeho dopady. Pravděpodobnost můžeme rozdělit do pěti stupňů. Rizika mohou být nepravděpodobná, málo pravděpodobná, pravděpodobná, velmi

pravděpodobná a nakonec téměř jistá. Předvídatelnost rizika je rovněž důležitá, jestliže lze riziko identifikovat a předvídat, je šance, že k nahodilosti nedojde. Další charakteristikou je míra ovlivnitelnosti. Rizika mohou být ovlivnitelná, částečně ovlivnitelná nebo neovlivnitelná. Interní rizika můžeme uvést jako příklad ovlivnitelných rizik, kdežto extérní rizika jsou neovlivnitelná. Pořadím působení rizik je myšlen vznik a následná odstranitelnost. Riziko může být malé, střední nebo velké a míra jeho přijatelnosti nezbytná, únosná a neúnosná.

Pojišťovnictví patří mezi finanční služby a mezi jeho úkoly patří zajištění pojistné ochrany občana, bezporuchový chod ekonomiky státu, spolupráce s bankovním sektorem na finančním trhu. Pojišťovnictví má dvě stránky, *etickou stránku*, jež je projevována v solidaritě ostatních pojištěných s postiženým- princip solidarity a *výdělečnou stránku*, jelikož se jedná o velmi prosperující odvětví pro podnikání. Pojišťovnictví je založeno na více odvětvích a to jak na ekonomii a financích, tak na pojistném právu a především na pojistné matematice.

2.2 Technické rezervy v neživotním pojištění

Technické rezervy jsou vytvářeny pojistitelem povinně podle zákona pro plnění závazků určených v pojistné smlouvě, které jsou pravděpodobné nebo jisté. Není známá pouze výše či okamžik vzniku těchto závazků. Technické rezervy řadíme mezi pasiva pojišťovny. Mezi technické rezervy patří, rezerva na nezasloužené pojistné, která odpovídá té části pojistného, jež se vztahuje k budoucím účetním obdobím. Rezerva na pojistná plnění je nejdůležitější technickou rezervou neživotního pojištění. Dalším typem je rezerva na prémie a slevy, jež je určena ke krytí nákladů na prémie a slevy poskytnuté pojistníkovi např. bonusy v havarijním pojištění. Posledním nejčastějším typem rezerv tvořených pojišťovnou jsou vyrovnávací rezervy. Vyrovnávací rezervy jsou určeny k vyrovnání vyšších pojistných plnění, které vznikly kolísáním škodního průběhu v důsledku událostí nezávislých na pojistiteli.

2.3 Charakteristika havarijního pojištění

V diplomové práci je pracováno se souborem dat týkajícího se havarijního pojištění, které je blíže popsáno v této podkapitole.

Havarijní pojištění je pojištěním dobrovolným, škodovým tj. pojistné plnění je omezeno rozsahem pojistného zájmu. Je řazeno mezi neživotní a majetkové pojištění. Předmětem havarijního pojištění, jinak kasko pojištění, jsou škody na motorových vozidlech. Jedná se o sdružené krytí rizik např. riziko havárie střetem, nárazem, živelní rizika, riziko vandalismu apod. Havarijní pojištění je často sjednáváno s výlukami a většinou se nevztahuje

na pojistné události vzniklé provozem vozidla, údržbou, opravou nebo nesprávnou obsluhou. Tímto druhem majetkového pojištění je kryta škoda na motorových vozidlech, kterou řidič neovlivnil, částečně ovlivnil nebo zcela ovlivnil. Havarijní pojištěním je kryto riziko havárie, ale je možnost rozšíření krytí i dalších rizik jako např. živelní rizika, odcizení vozidla, vandalismus aj. Často je uplatňováno rovněž pojištění asistenčních služeb. Pojištění se taktéž nevztahuje na škody vzniklé řízením vozidla pod vlivem návykových látek nebo bez oprávnění řízení vozidla. Pojistným plněním je myšlena časová cena vozidla snížená o cenu upotřebitelných zůstatků a to do výše horní hranice plnění či pojistné částky. Je možnost sjednání různých připojištění např. připojištění cestovních zavazadel, osob dopravovaných vozidlem aj. Pojistné je diferencováno podle obsahu motoru, stáří vozidla, regionu, spoluúčasti, účelu využívání atd. Typickým rysem havarijního pojištěni je spoluúčast. Spoluúčasti je myšleno podílení se pojištěného na úhradě škody. Pokud je v pojistné smlouvě sjednána spoluúčast, je placeno i nižší pojistné. Spoluúčast může být ve formě nějakého sjednaného procentního podílu na vzniklé škodě nebo formou excedentní spoluúčasti (maximální částka spoluúčasti na pojistném plnění).

U havarijního pojištění je často uplatňován systém bonusů a malusů. Bonusy jsou smluvně zaručené slevy na pojistném podle počtu bezškodných let, malusy jsou naopak přirážky k pojistnému podle uplatněných pojistných plnění v minulosti.

2.4 Risk management

Risk managementem dále RM myslíme řízení rizika v podniku. Ke vzniku risk managementu došlo pomocí vědeckých přístupů, kdy lidé chtěli zvládnout riziko.

Předmětem RM je zahrnutí projevů rizika vyplývajícího do rozhodování o hospodářských záležitostech v tržní ekonomice. Pod risk managementem si představíme analýzu ekonomické činnosti, jejímž cílem je minimalizace současného i budoucího rizika a omezování rozsahu přímých či nepřímých škod vzniklých jako důsledek rizikové situace. Cílem RM je dosažení bezpečné činnosti při vydání co nejmenších nákladů. Risk management je složen ze tří fází: identifikace rizika, zhodnocení rizika a zvládání rizika. První dvě fáze jsou označovány jako *analýzy rizika* a poslední fázi známe pod názvem *vlastní řízení rizik*.

Identifikaci rizika rozumíme zjištění rizikových faktorů, která mohou ohrožovat ekonomický subjekt. Subjekt je schopen se bránit pouze těm rizikům, které zná a může být na ně připravený. Správná identifikace je rozhodující částí risk managementu. Do těchto rizik

můžeme zahrnout riziko poškození majetku, fyzických ztrát nebo škod na zdraví. Dále zahrnujeme odpovědnost za škody, přerušení činnosti aj. Je důležité identifikovat jak vnitřní, tak i vnější rizika.

Druhou fází je **ocenění a kvantifikace rizik**. Přiřazujeme váhu jednotlivých rizikům a zjišťujeme, jaký dopad budou mít daná rizika na ekonomický subjekt. Do vyhodnocení rizika zahrnujeme zjištění pravděpodobnosti ztráty a její potenciální velikost, kdy zpravidla předpokládáme maximální možné škody.

Poslední fází risk managementu je **kontrola a financování rizik**. Jsou přijímána různá opatření, aby k pojistným událostem nedocházelo, a je vyvíjena snaha o eliminaci důsledků negativních nahodilostí. Nejdříve je rizikům předcházeno pomocí strategických a fyzických opatření. *Strategickými opatřeními* rozumíme změnu systému práce a do fyzických opatření patří aktivní, ale i pasivní prvky bezpečnosti, např. pořízení systému proti živelním rizikům, bezpečnější technologické postupy apod. *Preventivní ochrana* proti příčinám vzniku rizikových situací je tzv. *ofenzivní přístup*. Dalším typem je *eliminace* nepříznivých důsledků vzniklých při nahodilé události, jedná se o *defenzivní přístup k riziku*. U rizik, kterým nejde předejít u tzv. zbytkových rizik, je uvažováno o *finanční eliminaci*. Vzniklé škody je možno financovat z vlastních zdrojů, úvěrem nebo je možné rozdělit na více subjektů. Jedná se o krytí rizik z vlastních příjmů v podobě samopojištění, čili vytváření rezerv nebo z běžných příjmů. Z vlastních zdrojů jsou kryta rizika často se opakující a ta, kterými jsou způsobeny relativně malé ztráty. Hlavní výhodou využití vlastních zdrojů ke krytí škod je přímá angažovanost daného subjektu na ochraně proti riziku. Rizika v některých podnicích jsou kryta úvěrem. Tato varianta je méně vhodná, jelikož dochází k zanedbání prevence a není jistota, že úvěr v požadované výši subjekt dostane. Dalším typem krytí rizik je přenesení rizika na pojišťovnu, jak jsme již uvedli dříve.

Součástí risk managementu je *sledování rizik*, jejich pravidelná kontrola, následné vyhodnocování chování a identifikace změn. Je vypracováván tzv. *plán korekčních opatření*. V pravidelných časových intervalech je nutno prověřit efektivnost používaných metod prevence a systematicky sledovat změny v podniku, které mohou být příčinou vzniku dalších rizik. Dohled nad bezpečnostními opatřeními a efektivnosti pojištění je rovněž součástí risk managementu.

3 Charakteristika modelů četnosti

V dnešních časech, kdy na silnicích je mnoho motorových vozidel, často dochází k dopravním nehodám. V zájmu různých ekonomických subjektů je minimalizovat riziko dopravní nehody nebo alespoň přenést riziko na jiný subjekt. K tomu je používáno havarijní pojištění motorových vozidel. Pojistitelé ovšem na pojištění nechtějí prodělat a proto jsou tvořeny modely četnosti pojistných událostí a ty jsou následně používány k určení výše pojistného. Jednou z možností pro modelování četnosti je využití regresních modelů. Z toho důvodu je tato část práce zaměřena na metodické postupy, které budeme aplikovat na reálných datech v praktické části diplomové práce. Nejdříve je vysvětleno, co to je zobecněný lineární model. Jsou obecně popsány vzorce exponenciálního typu rozdělení a metody odhadu parametrů. V další částí všechny obecně zapsané vzorce jsou aplikovány na negativně binomické rozdělení a nakonec jsou popsány metody verifikace modelů. V celé kapitole je čerpáno z odborné literatury Hardman (2007) a Hilbe (2011). Také je vycházeno ze základní literatury Jong (2008), Ohlsson (2010) a Royston (2008).

3.1 Zobecněný lineární model

Metody regresní analýzy patří mezi jedny z nejčastěji používaných metod v oblasti matematické statistiky pro modelování a analýzu vícerozměrných dat, kde je kladen důraz na vztah mezi vysvětlující a vysvětlovanou proměnnou. Nejlepší odhad parametrů modelu by měl být získán regresní analýzou. Nejstarším, oblíbeným a stále hodně využívaným typem regresního modelu je klasický lineární regresní model. Odhady regresních koeficientů lineárních modelů je možno vypočítat pomocí poměrně jednoduchého vzorce. Jeho funkce z hlediska odhadovaných parametrů by měla být lineární. Ovšem nesmíme opomenout nevýhodu této metody, a to fakt, že při odvozování klasického lineárního regresního modelu jsou na zkoumaná data kladeny přísné předpoklady. V lineárním modelu by se neměla objevovat heteroskedasticita (rozptyl náhodné složky je variabilní), korelace mezi náhodnými složkami, autokorelace nebo to, že náhodné složky nepochází z normálního rozdělení. Pokud tyto podmínky nejsou při aplikaci splněny, můžeme dospět k úplně jiným výsledkům, popřípadě při výpočtech mohou vznikat komplikace. Mezi nesplněné předpoklady často patří předpoklad nezávislosti a normality dat. Nelineárním modelem může být exponenciální nebo mocninná funkce. V oblasti statistiky ale dochází k rozvoji, kterým je přineseno mnoho změn v odhadech. Jednou ze změn je možnost provádění složitých výpočtů, díky kterým se začaly ve větší míře aplikovat složitější regresní modely, k jejichž odhadu je nutné používat některé numerické metody. Mezi takovéto modely patří také zobecněné lineární regresní modely, jež

jsou určeny pro práci s daty s *rozdělením exponenciálního typu*. Do rodiny exponenciálních rozdělení patří kromě normálního rozdělení např. binomické, Poissonovo, gamma, negativně-binomické nebo exponenciální rozdělení.

Zobecněným lineárním modelem, dále jen GLM, je poskytován rámec pro vytváření jednotné třídy modelů. V těchto modelech je pracováno jak se spojitými, tak i kategorizovanými nezávislými proměnnými. Ve třídách jsou obsaženy regresní modely, obecné lineární modely, logistické a log-lineární modely.

Cílem GLM je nalézt model, kterým je *zmenšována* celková deviance (úměrnou rozdílu logaritmů funkcí věrohodnosti- mezi úplným modelem a nulovým modelem s jedním parametrem). Model je vytvářen postupně. Do modelu jsou zařazovány ty regresory, které nejvíce snižují devianci vzhledem k aktuálnímu modelu se zařazenými *k* parametry. Při vyhledávání vhodného vícenásobného regresního modelu může být použita metoda stepwise neboli postupné vkládání regresorů do modelu. Mezi typy nezávislých proměnných v odhadu jsou řazeny,

- kvalitativní faktory,
- iterace neboli (součiny) původních regresorů.

Pomocí GLM je možné odhadovat parametry složitých modelů. Zjednodušený postup konstrukce GLM můžeme napsat v pár krocích. Jako první je nutno určit *rozdělení pravděpodobnosti*, se kterým je následně pracováno. Druhým krokem je výběr *vhodné link funkce* pro dané rozdělení exponenciálního typu. Následně jsou vybrány vhodné *nezávislé proměnné* a jsou odhadovány jejich parametry v modelu. Na závěr je model *verifikován* a je *posouzena* jeho vhodnost.

Při odhadech je dobré začít nejdříve s jednorozměrnou analýzou. Jednorozměrnou analýzou je myšleno jednotlivé testování každé vysvětlující proměnné vůči závislé veličině. Po prvotním výběru významných proměnných hledáme nejlepší model s více proměnnými. Někdy je složité určit, které regresory (proměnné) je vhodné do modelu zahrnout. Pro výběr proměnných jsou použity tři způsoby. U *vzestupného* výběru se začíná pracovat s prázdným modelem a postupně jsou přidávány statisticky významné regresory, jedná se o tzv. *forward stepwise* metodu. Druhým typem je naopak *sestupný* výběr- *backward elimination*, kde se začíná od modelu se všemi proměnnými a postupně jsou odstraňovány statisticky nevýznamné proměnné. Posledním způsobem je, že jako první jsou odhadovány koeficienty

jednorozměrné analýzy, kde je zjištěno, které regresory jsou významné. Po zjištění významností koeficientů nezávislé proměnné jsou dosazeny do výsledného modelu. Pro nejlepší odhad modelu jsou v diplomové práci vybrány vhodné proměnné dle posledního postupu.

V reálu k odhadu je většinou k dispozici rozsáhlý soubor dat, který je možno rozdělit na dvě části. Na první částí dat je model odhadován a na té druhé je testován, čili je ověřena predikční schopnost a kvalita odhadovaného modelu.

Teorii zobecněných lineárních modelů představili v roce 1972 pan Nelder a Wedderburn. Zobecněné lineární modely jsou široce uplatňovány v různých oblastech medicíny, biologie či ekonomie a to především v pojišťovnictví. V pojišťovnictví je nejčastěji pracováno s Poissonovským a negativně binomickým rozdělením. Očekávaný počet pojistných událostí během vybraného období, čili škodní frekvenci je možné modelovat pomocí Poissonovské regrese nebo negativně-binomické regrese. Kdežto například výše škod pojistných událostí je odhadována pomocí gamma rozdělení.

3.2 Rozdělení exponenciálního typu

Základem zobecněných lineárních modelů je rozdělení exponenciálního typu, jimž je nahrazováno normální rozdělení pravděpodobnosti spojité náhodné veličiny v lineárním regresním modelu. Pomocí rozdělení exponenciálního typu jsme schopni modelovat chování střední hodnoty jak diskrétních, tak i spojitých náhodných veličin.

Dle Branda (2013) do exponenciálního typu rozdělení je zahrnováno například normální, Poissonovo, gama či alternativní rozdělení (speciální typ binomického rozdělení), ale známe-li i hodnotu některých parametrů, je započítáno taktéž rozdělení binomické, negativně binomické nebo Weibullovo.

Zobecněné lineární modely jsou tvořeny ze tří základních částí a to závislé proměnné, lineárního prediktoru a link funkce,

- závislá proměnná Y_i je ve tvaru náhodné veličiny pocházející z exponenciálního rozdělení, funkci můžeme zapsat v odvozeném tvaru jako,

$$f_y(y;\theta\ ,\phi) = \exp\left[\frac{y\theta - b(\theta)}{a(\phi)} + c(y,\phi)\right], \tag{3.1}$$

kde *a, b, c* jsou známé funkce, θ je neznámý standardní parametr nebo link funkce, který závisí na pozorování a ϕ je neznámý parametr disperze,

- lineární prediktor je ve tvaru,

$$\eta_i = \sum_{j=1}^{m} X_{ij}\beta_j = x_i^T \beta, \tag{3.2}$$

kde β_j jsou neznámé parametry a X_{ij} představují známé hodnoty regresorů,

- poslední základní částí zobecněného lineárního modelu je link funkce, jenž spojuje střední hodnotu závislé proměnné a lineární prediktor dle funkce,

$$E[Y_i] = \mu_i = g^{-1}(\eta_i). \tag{3.3}$$

3.3 Odhady parametrů

Při vytváření modelů většinou není známá přesná hodnota beta parametrů, a proto jsou parametry odhadovány na základě vzorku dostupných dat, jedná se tedy o přibližné odhady.

Rozlišujeme dva typy odhadů- bodový a intervalový,

- bodový odhad je takový, kde parametr je odhadován jedním číslem;
- v intervalovém odhadu je parametr aproximován intervalem, ve kterém se hodnota odhadovaného parametru s určitou pravděpodobností nachází.

V praktické části je pracováno s bodovými odhady, proto blíže popíšeme jejich vlastnosti, které by každý správný bodový odhad měl mít. Vlastnosti bodového odhadu jsou,

- nestrannost,
- vydatnost,
- konzistence,
- dostatečnost.

Mezi vlastnosti kvalitního bodového odhadu patří *nestrannost (nevychýlenost),* odhad je nestranný pouze tehdy, když jeho střední hodnota je rovna hledanému parametru,

$$E(\hat{\theta}) = \theta, \tag{3.4}$$

kde θ je hledaný parametr a $E(\hat{\theta})$ je střední hodnota odhadu.

Druhou vlastností je *vydatnost* (eficience). Máme-li dva nestranné odhady, tak je vybrán ten, jehož rozptyl je *nejmenší*. V níže popsaném vztahu by byl vybrán odhad číslo jedna,

$$\sigma^2(\hat{\theta}_1) < \sigma^2(\hat{\theta}_2), \tag{3.5}$$

kde $\sigma^2(\hat{\theta}_1)$ je nestranný odhad číslo jedna a druhý nestranný odhad je $\sigma^2(\hat{\theta}_2)$.

Konzistence je poslední vlastností správného odhadu, odhad je konzistentní tehdy, jeli přesnější s rostoucím výběrem. Dostatečným odhadem je myšlen odhad, ve kterém jsou obsaženy všechny informace o námi sledovaném parametru.

Pro odhady parametrů modelu jsou používány různé metody. Mezi klasické metody patří *metoda nejmenších čtverců*, *metoda momentů (MM)* nebo *metoda maximální věrohodnosti (ML)*. U negativně- binomického rozdělení (dále NB) by mělo platit, že střední hodnota modelu je menší než jeho rozptyl. Ve stručnosti je popsáno fungování metody nejmenších čtverců, metody momentů a následně se zaměříme na metodu maximální věrohodnosti, která je aplikována pro odhad modelu v praktické části diplomové práce.

3.3.1 Metoda nejmenších čtverců

Metoda nejmenších čtverců je nejvíce známou a taktéž nejvíce používanou metodou odhadu koeficientů. To, že je nejvíce používána neznamená, že musí být nejvhodnější. Je možné ji použít pouze pro odhad parametrů takových modelů, které jsou lineární. Metodu nejmenších čtverců není možné použít u nelineární regrese. Metoda nejmenších čtverců je používána k nalezení takového řešení, kdy součet druhých mocnin odchylek odhadu by měl být *minimální*, čili součet čtverců chyb by měl být nejmenší. Rezidua neboli chyby odhadu jsou vypočteny jako rozdíl mezi napozorovanou (empirickou) a odhadnutou (teoretickou) hodnotou a je možné je chápat jako chyby, kterých je v daném bodě odhadu dopouštěno. Minimalizaci reziduálních čtverců vypočítáme následovně,

$$S_i^{\,2} = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 \rightarrow \min , \tag{3.6}$$

kde Y_i je napozorována hodnota a $\hat{Y}_i$ je odhadována hodnota.

Místo odhadnuté hodnoty dosadíme konkrétní tvar regresní přímky. Celý výraz je parciálně derivován dle jednotlivých parametrů a postavíme rovno nule. Následně získáme soustavu rovnic, kterou vyřešíme a obdržíme hodnoty odhadovaných parametrů. Zda se jedná o minimum nebo maximum zjistíme tím, že vypočítáme druhé derivace rovnic. Vyjde-li druhá derivace záporná, jedná se o minimum.

Metoda nejmenších čtverců by měla mít tyto předpoklady,

- odhad by měl být lineární v parametrech;
- veličiny by měly být náhodné;

- náhodné chyby odhadu jsou homoskedasticitní- s konstantním rozptylem v čase a mají nulovou střední hodnotu;
- proměnné modelu by neměly být na sobě závislé;
- v modelu by se neměla vyskytovat multikolinearita a ani autokorelace;
- odhadované chyby by měly vycházet z normálního rozdělení.

3.3.2 Metoda momentů

Metoda momentů je jednoduchá metoda sloužící k odhadům parametrů známých rozdělení. Jsou *porovnávány* výběrové momenty s momenty teoretickými předpokládaného rozdělení. Mezi vlastnosti odhadů metody momentů patří, že odhady jsou eficientní, čili odhad má co nejmenší rozptyl, dále odhad by měl být asymptoticky normální, vychýlený a ne vždy robustní.

Jestliže náhodný výběr je dán hodnotami Y_i, pak k-tý výběrový moment je dle vzorce,

$$M_k = \frac{1}{n}\sum_{i=1}^{n} Y_i^k, \tag{3.7}$$

a centrální moment je dán vzorcem,

$$M_k = \frac{1}{n}\sum_{i=1}^{n} (Y_i - \bar{Y})^k, \tag{3.8}$$

kde $\bar{Y}$ je výběrový průměr. Teoretické momenty jsou následující, počáteční moment je ve tvaru,

$$\mu'_k = \sum_{i=1}^{n} Y_i^k p(Y_i), \tag{3.9}$$

a centrální moment je,

$$\mu_k = \sum_{i=1}^{n} (Y_i - \mu'_1)^k p(Y_i). \tag{3.10}$$

Má-li rozdělení s hustotou pravděpodobnosti $f(Y)$ počet r neznámých parametrů a soustava rovnic je dána vzorcem,

$$M_k = \mu_k, \tag{3.11}$$

kde $k \in N$ a soustava rovnic má jediné řešení, metodou momentů jsou dány odhady r parametrů.

3.3.3 Metoda maximální věrohodnosti

Metoda maximální věrohodnosti jinak MLE je jednou z centrálních metod statistiky. Jejím cílem je odhad neznámých veličin v závislosti na pozorovaných datech. Odhad neznámých veličin je rozdělen na formulaci pravděpodobnostního modelu a srovnání odhadovaného modelu,

- formulace pravděpodobnostního modelu, jež slouží k popisu dané situace;
- srovnání odhadovaného modelu se skutečností vycházející z reálných dat.

Principem metody maximální věrohodnosti je najít odhad parametru θ (případně vektoru parametrů), kterým je maximalizována pravděpodobnost, že pozorované hodnoty pocházejí z předpokládaného rozdělení pravděpodobnosti. Snažíme se najít takovou hodnotu, pro niž je pravděpodobnost, že pozorované hodnoty pocházející z předpokládaného rozdělení jsou maximální. Odhadem se tedy snažíme maximálně přizpůsobit pozorovaným datům, když je připuštěno, že daty je představen jediný zdroj informací neznámého parametru. Mezi vlastnosti odhadu metody maximální věrohodnosti patří, konzistentnost, eficience a odhad by měl být asymptoticky normální.

Pozorovaná data jsou brána jako určitý soubor náhodných veličin Y_i se stejným rozdělením a neznámou distribuční funkcí f_θ. Existuje taková hodnota θ_0, že $f_\theta = g_{\theta_0}$. Hodnota je neznámá a cílem odhadu $\hat{\theta}$ je se této hodnotě co nejvíce přiblížit. Obecný zápis pro MLE odhad je uveden v následujících krocích.

Distribuční funkci stejně rozdělených a náhodných hodnot můžeme popsat následovně,

$$f(Y_1,...,Y_n|\theta) = \prod_{i=1}^{N} f(Y_i|\theta). \tag{3.12}$$

Střední hodnota u exponenciální formy je ve tvaru,

$$E(Y) = (\theta_i), \tag{3.13}$$

a rozptyl obecně můžeme napsat jako,

$$\text{var}(Y) = (\theta_i). \tag{3.14}$$

Chceme-li odhadovat hodnoty θ, rovnice pro odhad je,

$$L(\theta,\phi|Y_1,...,Y_n)=\prod_{i=1}^{N} f(Y_i|\theta,\phi)\,. \tag{3.15}$$

Tato rovnice je funkcí věrohodnosti. Pro komplikovaný výpočet součinu je funkce věrohodnosti upravována- logaritmována a je dána vzorcem,

$$\log L(\theta,\phi|Y_1,...,Y_n)=\sum_{i=1}^{N}\log f(Y_i|\theta,\phi)\,. \tag{3.16}$$

Zlogaritmovanou funkci věrohodnosti pro exponenciální typ rozdělení lze zapsat ve tvaru,

$$\log L=\sum_{i=1}^{n}\left\{\frac{y_i\theta_i-b(\theta_i)}{a(\phi)}+c(y_i,\phi)\right\}, \tag{3.17}$$

kde $c(y_i,\phi)$ je tzv. normalizační konstanta a jako jediná nezávisí na datech,

Se součtem se lépe pracuje, proto tato modifikace zlogaritmované funkce věrohodnosti je častěji používaná. Jestliže platí,

$$L(\theta,\phi|Y_1,...,Y_n)\le L(\hat{\theta},\phi|Y_1,...,Y_n)\,, \tag{3.18}$$

tak $\hat{\theta}$ je maximální věrohodný odhad.

První parciální derivací zlogaritmované funkce věrohodnosti je tzv. *gradient*, derivaci lze získat dle vzorce,

$$\partial\log L=\frac{\partial\log L}{\partial\beta}\,, \tag{3.19}$$

kde β je odhadnutý koeficient.

Druhou parciální derivaci nazýváme *Hessian* (Hessian matice). Derivaci lze spočítat následovně,

$$\partial^2\log L=\frac{\partial^2\log L}{\partial\beta\partial\beta'}\,. \tag{3.20}$$

Negativní, inverzní matice k Hessianu je tzv. *variančně- kovarianční matice*. Standartní chyby odhadu jsou založeny na elementech na diagonále této matice, kterou je možno pojmenovat jako informační matice. Zobecněné lineární modely jsou odhadovány

pomocí *Newton- Raphson* metody nebo metodou *IRLS*. Newton- Raphton algoritmem lze získat *pozorovanou informační matici* (OIM), naopak metodou iterativně vážených nejmenších čtverců (IRLS) získáme *očekávanou informační matici* (EIM).

3.4 Negativní binomická regrese

Negativní binomické rozdělení je řazeno mezi jedno ze základních diskrétních rozdělení pravděpodobnosti. Toto rozdělení je často využíváno na příklad v pojišťovnictví nebo v lékařských, psychologických či biologických oborech. Negativní binomická regrese je zvláštním typem Poissonovské regrese. V této podkapitole je vycházeno z odborných článků Valecký (2012a, 2012b) a Valecký (2015).

3.4.1 Rozdělení pravděpodobnosti

V této podkapitole jsou popsány základní veličiny z negativně binomického rozdělení pravděpodobnosti.

Negativní binomické rozdělení vychází z Bernoulliovské posloupnosti nezávislých postupů, kdy pravděpodobnost úspěchů π v jednotlivém pokusu je $\pi \in (0;1)$ a počet neúspěchů je $1-\pi$. Negativní binomické rozdělení je takový exponencionálním typem rozdělení pravděpodobnosti, kdy Y je náhodná veličina a udává počet neúspěchů předcházejících k-tému úspěchu $k \in \{1,2,...\}$. Parametry modelu jsou π a k. Pravděpodobnostní funkci můžeme vyjádřit takto,

$$f(y;k;\pi) = \binom{y+k-1}{y} \cdot (1-\pi)^{y} \cdot \pi^{k}. \tag{3.21}$$

V diplomové práci předpokládáme, že pro parametry NB rozdělení platí $k > 0$ a $0 < \pi < 1$. Střední hodnota a rozptyl negativně binomického rozdělení jsou popsány vzorci,

$$\mu = E(Y) = k\frac{1-\pi}{\pi}, \tag{3.22}$$

kde μ je střední hodnota. Rozptyl je popsán následovně,

$$\sigma^2 = \mathrm{var}(Y) = k\frac{1-\pi}{\pi^2}, \tag{3.23}$$

kde σ^2 je rozptyl. Pravděpodobnostní funkci si můžeme upravit zavedením proměnné p, kdy $\pi = \frac{1}{1-p}$. Pravděpodobnostní funkce negativně binomického rozdělení je ve tvaru,

$$f(y;p;k)=\binom{-k}{y}\cdot p^{y}\cdot(1-p)^{-k-y}. \tag{3.24}$$

Z důvodu záporného $-k$ je rozdělení nazýváno negativně binomické.

V následující vzorcích je použit jiný typ zápisu negativně binomického modelu. Negativně binomický model je takový model, kdy parametr α je větší jak nula, proto odhady parametrů jsou rozdílné od odhadů parametrů v Poissonovské regresi (koeficient α=0). Model NB2 je nekanonický typ negativního binomického modelu. Nekanonický (standardní) typ negativně binomické regrese je specifickým druhem Poissonovské regrese. Model je také nazýván NB2 modelem vzhledem k povaze kvadratické funkce rozptylu. V praktické části diplomové práce je odhadován model s NB2 rozdělením.

Střední hodnota je dána vzorcem,

$$E(Y)=\mu_i, \tag{3.25}$$

a rozptylová funkce je ve tvaru,

$$\text{var}(Y)=\mu_i+\alpha\cdot\mu_i^{2}. \tag{3.26}$$

Funkce maximální věrohodnosti, kde náhodné veličiny jsou z negativně binomického rozdělení, je dána vzorcem,

$$L(k,\pi)=\prod_{i=1}^{n} f(y_i;k,\pi), \tag{3.27}$$

Logaritmicky zápis funkce věrohodnosti pro n náhodných veličin je,

$$\log L(k,\pi)=\sum_{i=1}^{n}\ln(\Gamma(y_i+k))-\sum_{i=1}^{n}\ln(k_i!)-n\cdot\ln\cdot(\Gamma(k))+\sum_{i=1}^{n}y_i\cdot\ln(\pi)+n\cdot k\cdot\ln(1-\pi). \tag{3.28}$$

Tento zápis funkce je vhodný pouze v případech, kdy odhadované parametry jsou pro všechny náhodné veličiny stejné, jeli ale některý z odhadovaných parametrů pro každou náhodnou veličinu jiný, musíme použít jiný zápis MLE funkce a to,

$$\log L=(y_i;\mu_i,\alpha)=\left\{\begin{array}{l}\sum_{i=1}^{n}y_i\cdot\ln(\frac{\alpha\cdot\mu_i}{1+\alpha\cdot\mu_i})-\frac{1}{\alpha}\ln(1+\alpha\cdot\mu_i)+\ln\Gamma(y_i+\frac{1}{\alpha})-\ldots\\ \ldots-\ln\Gamma(y_i+1)-\ln\Gamma(\frac{1}{\alpha})\end{array}\right\}, \tag{3.29}$$

kde $k = \frac{1}{\alpha}$ a $\pi = \frac{1}{(1+\alpha \cdot \mu_i)}$.

Deviance u negativně binomického rozdělení je ve tvaru,

$$D = 2\sum_{i=1}^{n}\left\{ y_i \ln\left(\frac{y_i}{\mu_i}\right) - \left(y_i - \frac{1}{\alpha}\right)\ln\left(\frac{1+\alpha y_i}{1+\alpha\mu_i}\right)\right\}. \tag{3.30}$$

3.4.2 Link funkce

Link funkcí je představena transformována střední hodnota, kde link funkce z negativně binomického rozdělení je,

$$g(\mu_i) = \theta_i = \ln(\mu_i) = x_i\beta, \tag{3.31}$$

kde x_i jsou vysvětlující proměnné a β je odhadované koeficienty.

Tato funkce je nekanonická a inverzní link funkce je ve tvaru,

$$g^{-1}(\theta_i) = \mu_i = \exp(x_i\beta). \tag{3.32}$$

Model s těmito tvary nazýváme NB2 model. Následně log-likelihood funkce, která zahrnuje inverzní link funkci je v následující podobě,

$$\log L = (y_i;\beta,\alpha) = \left\{ \begin{array}{l} \sum_{i=1}^{n} y_i \cdot \ln(\frac{\alpha \cdot \exp(x_i\beta)}{1+\alpha\cdot\exp(x_i\beta)}) - \frac{1}{\alpha}\ln(1+\alpha\cdot\exp(x_i\beta)) + \ldots \\ \ldots + \ln\Gamma(y_i + \frac{1}{\alpha}) - \ln\Gamma(y_i+1) - \ln\Gamma(\frac{1}{\alpha}) \end{array} \right\}. \tag{3.33}$$

3.4.3 Expozice

V pojišťovnictví je často pracováno s modely, které mají určitou expozici v riziku např. počet rizik, délka platnosti smlouvy aj. Proto je nutno provést korekci modelu a tuto expozici zavést do lineárního prediktoru např. dle vzorce,

$$\mu_i = \exp[(x_i \cdot \beta) + \ln(t_i)], \tag{3.34}$$

kde t_i je délka platnosti jednotlivé smlouvy.

3.4.4 Derivace funkce maximální věrohodnosti

První parciální derivací vzhledem parametru beta získáme tzv. *gradient*,

$$\frac{\partial \log L}{\partial \beta} = \sum_{i=1}^{n} \frac{x_i(y_i - \mu_i)}{1 + \alpha\mu_i} \ , \tag{3.35}$$

pokud parametr α není v modelu konstantou, první parciální derivace vzhledem parametru α je následující,

$$\frac{\partial \log L}{\partial \alpha} = \sum_{i=1}^{n} \left[\frac{1}{\alpha^2} \left(\ln(1 + \alpha\mu_i) + \frac{\alpha(y_i - \mu_i)}{1 + \alpha\mu_i} \right) + \psi(y_i + \frac{1}{\alpha}) - \psi\left(\frac{1}{\alpha}\right) \right] . \tag{3.36}$$

Pro odhad parametru parciální derivaci máme postavit rovno nule.

Rovněž je nutno rozlišit očekávanou a pozorovanou informační matice. Pozorovaná informační matice je brána jako záporný Hessian. Model standardních odchylek u nekanonických modelů je druhou odmocninou na diagonále negativní inverzní Hessian matice. Obecně platí, že nemáme-li malý počet pozorování, standartní chyby odhadu očekávané informační matice se blíží k chybám pozorované informační matice. Poslední studie dokázaly, že u nekanonických modelů pozorované standartní odchylky jsou asymptoticky méně zkreslené než očekávané standartní chyby.

Existují dva způsoby odhadu parametrů a to pomocí metody iterativních vážených nejmenších čtverců, dále jen IRLS nebo pomocí Newton- Raphsonova algoritmu. Metodou Newton- Raphson je získána pozorovaná informační matice ve tvaru,

$$\hat{V}_{OH} = \left\{ (-\frac{\partial^2 \log L}{\partial \beta_j \partial \beta_k}) \right\}^{-1} , \tag{3.37}$$

kde elementy matice $-\frac{\partial^2 \log L}{\partial \beta_j \partial \beta_k}$ vypočítáme dle vzorce,

$$\frac{\partial^2 \log L}{\partial \beta_j \partial \beta_k} = \left\{ -\sum_{i=1}^{n} \frac{1}{a(\phi)} \left[\begin{array}{l} \frac{1}{V(\mu_i)^2} \left(\frac{\partial \mu}{\partial \eta} \right)^2_i - \ldots \\ \ldots - (\mu_i - y_i) \cdot \left\{ \frac{1}{V(\mu_i)^2} \left(\frac{\partial \mu}{\partial \eta} \right)^2_i \frac{\partial V(\mu_i)}{\partial \mu_i} - \frac{1}{V(\mu_i)} \left(\frac{\partial^2 \mu}{\partial \eta^2} \right)_i \right\} \end{array} \right] x_{ji} x_{ki} \right\} \tag{3.38}$$

Zatímco tradiční metodou IRLS je použita očekávaná matice, která je dána vzorcem,

$$\hat{V}_{EH} = \left\{ -E\left(\frac{\partial^2 \log L}{\partial \beta_j \partial \beta_k} \right) \right\}^{-1} , \tag{3.39}$$

kde $\frac{\partial^2 \log L}{\partial \beta_j \partial \beta_k}$ spočítáme následovně,

$$E\left(\frac{\partial^2 \log L}{\partial \beta_j \partial \beta_k}\right) = -\sum_{i=1}^{n} \frac{1}{a(\phi)} \frac{1}{V(\mu_i)} \left(\frac{\partial \mu}{\partial \eta}\right)^2_i x_{ij} x_{ki}. \tag{3.40}$$

Tyto metody slouží k odhadu standardních odchylek v modelu.

3.5 Verifikace

Verifikaci jsou myšleny některé ověřovací metody sloužící pro určení vhodnosti daného modelu. Ve více případech jsou platné i pro regresní modely s různým rozdělením. Jako první jsou popsána rezidua, následně je uveden základní test (R^2), deviance a ukazatel věrohodnosti. Nakonec jsou popsány informační kriteriální statistiky a to AIC a BIC.

3.5.1 Rezidua

Tradiční forma NB2 rozdělení má přirozenou log link funkci, i když není zapsaná v kanonické podobě. Rezidua je možno využít k posouzení kvality modelu, detekci odlehlých pozorování a ověření předpokladu o rozptylu. V práci jsou zmíněna základní rezidua a to Pearsnova a rezidua deviance. Základní rezidua jsou definována jako rozdíl mezi pozorovanými a predikovanými proměnnými dle vzorce,

$$r_i = y_i - \hat{y}_i = y_i - \mu_i, \tag{3.41}$$

kde $\hat{y}_i$ je predikovaná hodnota.

Dalším typem reziduí jsou rezidua deviance a Pearsnova rezidua. Rezidua deviance jsou dána vzorcem,

$$r_i^{\,d} = \operatorname{sgn}(y_i - \mu_i) \cdot \sqrt{d(y_i \cdot \mu_i)}, \tag{3.42}$$

kde $\sqrt{d(y_i \cdot \mu_i)} = 2\left[y_i \cdot g(y_i) - b\left(g\left(y_i\right)\right) - y_i \cdot g\left(\mu_i\right) + b\left(g\left(\mu_i\right)\right)\right]$ je distanční funkce, kterou je představena odlehlost od odhadované střední hodnoty μ_i k y_i.

Rezidua deviance jsou používána k identifikaci odlehlých pozorování.

Pearsonova rezidua lze vypočítat dle vzorce,

$$r_i^{\,p} = \frac{y_i - \mu_i}{\sqrt{V\left(\mu_i\right)}}, \tag{3.43}$$

kde $V(\mu_i)$ je rozptyl.

Rovněž existují standardizovaná rezidua, kdy Pearsonova rezidua nebo rezidua deviance jsou vydělena $\sqrt{1-h_{ii}}$, kde h_{ii} je i-tý diagonální prvek matice a $h_{ii} = stdp^2 \cdot V(\mu_i)$, kdy *stdp* je standartní chyba odhadu.

Diagonální prvek matice je popsán následovně,

$$h_{ii} = V^{\frac{1}{2}} \cdot X \cdot (X'VX)^{-1} \cdot X'V\frac{1}{2}, \tag{3.44}$$

kde V je matice s prvky na diagonále

$$v_{ii} = \left(\frac{\partial \mu_i}{\partial x_i b}\right)^2 V(\mu_i)^{-1}. \tag{3.45}$$

Hodnoty standardizovaných Pearsonových reziduí je vhodné znázornit graficky. Pokud se nachází hodně hodnot mimo interval <-2;2>, tak model není vhodný.

Je důležité ale zdůraznit, že při modelování četnosti pojistných škod rezidua nejsou tak užitečným nástrojem jak u normální lineární regrese, neboť nemají asymptoticky normální rozdělení. Ale všechna výše uvedena rezidua získávají přibližně normální rozdělení za předpokladu, že je pracováno s modelem seskupených dat.

3.5.2 Statistika vhodnosti modelu

Základním ukazatelem, zda je daný model významný, je *p-value* neboli významnost. Je-li hodnota nižší než 0,05, pak na hladině významnosti 5% je zamítnuta vybraná nulová hypotéza a naopak je přijata hypotéza alternativní. Hodnocení ukazatelem p-value je významné především u jednorozměrné analýzy.

Dalšími ukazateli významnosti modelu jsou pseudo R^2 statistika, devianční statistika a ukazatel věrohodnosti. Prvními ukazateli je testován model a dalšími ukazateli je srovnáváno, který typ rozdělení pro daný model je vhodnější. V diplomové práci je počítáno pouze s negativně binomickou regresí, protože již bylo dokázáno, že pro modelování četnosti je vhodnější NB regrese než Poissonovská. Z toho důvodu ukazatelem věrohodnosti budou srovnávány modely s různými proměnnými.

Ukazatel pseudo- R^2 má podobný význam jako koeficient determinace, čím je jeho hodnota vyšší, tím je model vhodnější a vypočítáme jej jako,

$$R^2{}_P = 1 - L_F / L_I , \tag{3.46}$$

kde L_F je hodnota funkce maximální věrohodnosti celého modelu a L_I je hodnota funkce věrohodnosti modelu, který zahrnuje pouze konstantu.

Deviance

Devianční statistika je používaná v zobecněných lineárních modelech často a je zobrazována automaticky u většiny výstupu ze statistických softwaru. Deviance je vyjádřena vzorcem,

$$D = 2\sum_{i=1}^{n} \{\log L(y_i;y_i) - \log L(\mu_i;y_i)\}, \tag{3.47}$$

kde $\log L(y_i;y_i)$ je taková log-likelihood funkce, kde každé hodnotě μ_i je přiřazena hodnota y_i a $\log L(\mu_i;y_i)$ je logaritmovaná funkce maximální věrohodnosti odhadovaného modelu.

Deviance má význam u srovnávání dvou modelů s různým rozdělením pravděpodobnosti. Lepší je ten model, jehož hodnota deviance je nižší. Dle Hilbe (2011, s. 67) tato statistika již není v dnešní době používána v takovém rozsahu pro stanovení vhodnosti modelu, kdy často modely, které po testování vyšly jako vhodné, v reálu vhodné nejsou.

Ukazatel věrohodnosti

Ukazatel je často používán při srovnávání modelů s různým rozdělením pravděpodobnosti. Pro tento test je nutno vypočítat odhady parametrů jak pro úplný model, tak i pro model redukovaný, jenž je speciálním tvarem plného modelu. Testová statistika je dána vzorcem,

$$LR = -2(\log L_I - \log L_F) . \tag{3.48}$$

Věrohodnostní statistika má při velkých výběrech prostřednictvím Pearsonovy statistiky χ^2 - rozdělení s počtem stupňů volnosti, který je roven rozdílu počtu parametrů testovaných modelů,

$$LR \sim \chi^2{}_{kf-ki}(0,1-\alpha), \tag{3.49}$$

kde $_{kf-ki}$ jsou stupně volnosti.

Má-li parametr v plném modelu významný vliv na vysvětlovanou proměnnou, je hodnota LR testu vysoká. Pokud ale je vliv parametrů v plném modelu malý, zanedbatelný, jsou hodnoty obou modelů téměř stejné. Je-li hodnota pravděpodobnosti menší než 0,05 na hladině významnosti 5%, koeficienty jsou statisticky významné. Test poměrem věrohodnosti je alternativou pro *F*-test u lineárního regresního odhadu. Ukazatel *LR* je vhodný, jestli uvažujeme o přidání nějakého regresoru do modelu.

Waldův test

Test je používán pro testování významnosti jednotlivých parametrů. Ve Waldově testu stačí odhadnout parametry v jednom modelu a ty následovně testovat. Podmínkou je rozsáhly soubor dat. Nulová hypotéza je stanovena jako,

$$H_0 : \hat{\beta} = 0, \tag{3.50}$$

a alternativní hypotéza je ve znění,

$$H_A : \hat{\beta} \neq 0, \tag{3.51}$$

kde $\hat{\beta}$ je odhad testovaného parametru pomocí MLE.

Koeficient je statisticky nevýznamný, jeli jeho hodnota rovná nule. Naopak pokud je koeficient různý od nuly, proměnná je statisticky významná a měl by být v modelu použita.

Testová statistika Waldova testu má χ^2 rozdělení s jedním stupněm volnosti a je dána vzorcem,

$$W = \left(\frac{\hat{\beta}}{s_{\hat{\beta}}} \right)^2, \tag{3.52}$$

kde $s_{\hat{\beta}}$ je standardní odchylka odhadu daného parametru.

Pokud je p-value nižší než 0,05, je zamítnuta nulová hypotéza a odhadovaný parametr je na hladině významnosti 5% statistický významný a může být v modelu ponechán. Waldova statistika je alternativou *t-testu* v lineární regresi. Test je vhodný pro určení významnosti parametru, ale má jistá omezení, proto je nutné použít i další možnosti verifikace.

Velmi podobných výsledků při velkém počtu pozorování můžeme získat jak u věrohodnostní statistiky, tak u Waldova testu. Při malém výběru se ale výsledky mohou

hodně lišit. Většinou se dává přednost výpočtu testu poměrem věrohodností, ale Waldův test je výpočetně jednodušší.

3.5.3 Informační kritéria AIC a BIC

Informačními kritérii je měřena kvalita statistického modelu pro určitý datový soubor. Každé kritérium zahrnuje určité alternativní parametrizace. Preferovány jsou modely s nižšími hodnotami informačních matic, kterými je indikována lepší „vhodnost“ modelu. Při srovnání modelů pomocí informačních kritérií je zohledňována nejen hodnota funkce věrohodnosti, ale i počet parametrů.

Akaikeho informační kritérium dále AIC vypočítáme jako,

$$AIC = \frac{-2(\log L - k)}{n}, \tag{3.53}$$

kde $\log L$ je log-likelihood model, neboli logaritmus funkce věrohodnosti, kde k je počet prediktorů a n je počet pozorování. Druhý typ výpočtu je,

$$AIC = -2(\log L - k). \tag{3.54}$$

Před použitím hodnoty AIC pro srovnání je třeba zjistit, který vzorec výpočtu je používán v statistickém softwaru.

Druhým nejvíce používaným statistickým kritériem je Bayesovo informační kritérium dále BIC. První vzorec BIC je založen na statistice deviance a je ve tvaru,

$$BIC_R = D - (df)\ln(n), \tag{3.55}$$

kde D je model statistiky deviance a df jsou stupně volnosti modelu, kdy od pozorovaných hodnot odečteme počet prediktorů modelu včetně konstanty. Ve výstupu ze STATA 11.0 při použití příkazu pro zobecněný lineární model je získán tento typ ukazatele. V následující Tab. 3.1 jsou zobrazeny stupně preference při srovnávání vhodnosti dvou modelů pomocí *BIC* statistiky.

Tab. 3.1 Hodnoty srovnání ukazatelů BIC

Rozdíl	Stupně preference
0-2	Slabý
2-8	Pozitivní
6-10	Silný
>10	Velmi silný

Zdroj: (Hilbe, 2011)

U většiny ostatních příkazů ze softwaru STATA 11.0 je výstupem tato parametrizace BIC,

$$BIC_L = -2\log L + k \cdot \ln(n), \tag{3.56}$$

kde k je počet prediktorů(koeficientů) včetně konstanty. Nesmíme opomenout, že hodnoty těchto dvou různých parametrizací se liší a proto je důležité vědět, který typ výpočtu v softwaru je použit.

3.6 Vysvětlující proměnné

V diplomové práci v souboru dat jsou obsažené jak spojité, tak i kategorické veličiny. Typy vysvětlujících proměnných jsou důležité při tvorbě modelu, odhadu parametrů a hodnocení kvality modelu. Nezávislé proměnné by se neměly opakovat. Tento problém není u spojitých veličin, ale u kategorických ano. Data musí být rozdělena do tabulky a četnost modelována tříděním proměnných. Je-li vysvětlující proměnná kategorickou veličinou, při odhadu parametrů se vychází z relativních četností výskytu sledovaného jevu pro danou kombinaci hodnot nezávislých proměnných rozdělených do tabulek. Kategorizace spojitých proměnných by měla být prováděna s opatrností, jelikož se jedná o ztrátu informace, kterou může být zvýšena variabilita modelu a to by mohlo vést ke zkreslení výsledků.

Pokud u kategorických veličin nám vyjde některá z kategorií nevýznamná, nemusíme hned vynechávat danou kategorii nebo proměnnou, ale nejdříve kategorickou veličinu podrobíme Waldovu testu dle vzorce (3.52), zda jako celek je významná. Pokud jako celek vyjde významná, můžeme ji ponechat v námi odhadovaném modelu.

4 Odhad a vyhodnocení regresního modelu

Aplikovaná část práce je rozdělena do několika kapitol. V první kapitole jsou krátce popsána data, na kterých je modelována četnost pojistných událostí. Následně se zaměříme na jednofaktorovou analýzu, abychom zjistili, které vysvětlující proměnné jsou významné a mají vliv na četnost pojistných událostí v modelu. V aplikované části je analyzován vliv jednotlivých regresorů na škodní frekvenci. Předpokládáme, že náhodné veličiny pocházejí z negativně binomického rozdělení a jednotlivé parametry jsou odhadovány metodou maximální věrohodnosti. Po určení významnosti jednotlivých prediktorů, je vytvořen finální model, který je následně verifikován, případně je upraven. V další části je srovnán „plný" model s modelem upraveným a pomocí testu poměrem věrohodnosti je určeno, zda je lepší model „plný" nebo upravený. Nesmíme zapomenout na exposure- duraci, jak dlouho havarijní pojištění trvalo, jestli celý rok nebo jen poměrnou část roku. Tato veličina má vliv na frekvenci škod. Kdyby nebyla časová expozice zahrnutá v modelu, bylo by modelováno pouze počet škod v havarijním pojištění.

Cílem práce je modelování četnosti škod v havarijním pojištění a také vliv kategorických veličin na škodní frekvenci, tak v druhé části práce všechny spojité proměnné, které jsme použili k odhadování modelu, jsou přeměněny na kategorické veličiny. Znova je odhadován model negativní binomické regrese s kategorizovanými veličinami. A nakonec je porovnán s modelem, ve kterém byly použity jak kategorické, tak i spojité veličiny.

4.1 Data

Účelem této práce je aplikace metod uvedených v předchozích kapitolách na reálném souboru dat. Cílem práce je modelování četnosti pojistných škod v havarijním pojištění. Pomocí metody věrohodnosti odhadujeme nejvhodnější parametry pro model. K odhadu parametrů jednotlivých proměnných je použit statistický software STATA 11.0. Základním souborem dat, ze kterého vycházíme v praktické části, je soubor proměnných s náhodným výběr četnosti škod v havarijním pojištění v letech 2005 až 2010. Celkem se jedná o 233 879 pozorování. V práci jsou použity zkratky proměnných. Agecar je stáří vozidla, ageman je věk řidiče, proměnnou price je představena kupní cena vozidla. Dalším známým parametrem je, zda se jedná o právnickou či fyzickou osobu, jež je reprezentováno zkratkou company. Pohlaví pojištěného je gender, dále doba trvání pojištění v roce, neboli časová expozice je v práci zastupována zkratkou duration. Ale rovněž informace týkající se vozidla a to, jaký má výkon (kw), objem motoru (volume) nebo typ paliva, na které jezdí- fuel atd. Proměnnou district jsou myšleny kraje v České republice, ve kterých došlo k pojistným událostem.

Jedná se o různé kategorie veličin od spojitých přes kategorické až po binární. Binární veličina je zvláštním typem veličiny kategorické. V následující Tab. 4.1 jsou popsány základní statistické charakteristiky u veličin, které jsou následně podrobeny jednofaktorové analýze,

Tab. 4.1 Základní charakteristika proměnných

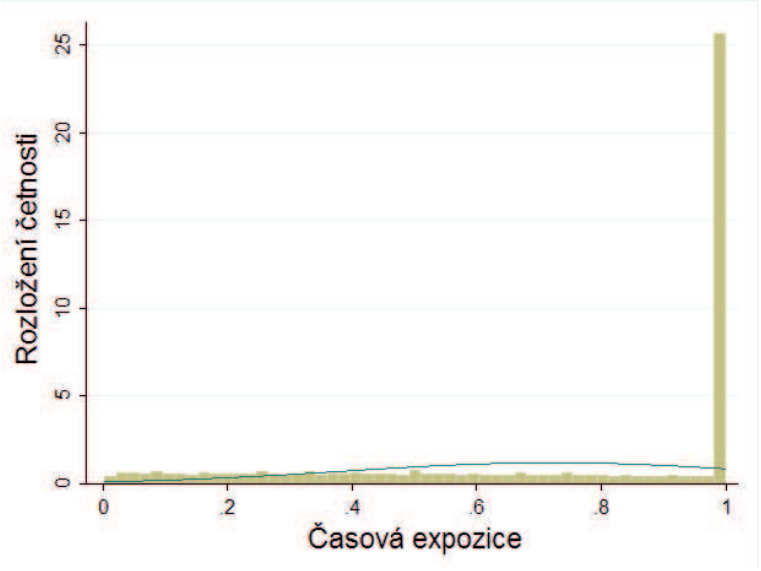
Proměnná	Min	Max	Střední hodnota	Sm.odchylka	Šikmost	Špičatost
Company	0	1	0,688	0,463	-0,812	1,660
Gender	0	1	0,224	0,417	1,325	2,757
Fuel	0	3	0,252	0,439	1,268	3,258
Duration	0,003	1	0,724	0,332	-0,761	2,061
Excess	5	30	6,503	2,988	2,662	13,609
Kw	29	456	69,033	30,411	2,788	17,832
Volume	599	7011	1597,643	476,095	2,564	16,073
District	0	13	8,280	3,754	-0,723	2,445
Price	1999	8186265	392602,8	297460,9	4,825	54,640
Ageman	18	108	48,64	13,264	0,109	2,246
Agecar	0	44	4,579	3,192	0,608	3,641

Zdroj: vlastní zpracování v STATA 11.0

Všechny vysvětlující proměnné uvedené v Tab. 4.1 jsou špičatější než proměnné s normálním rozdělením pravděpodobnosti. Většina z nich je rovněž pravostranně asymetrická, kdy hodnota šikmosti je kladná. Veličinou subjekt je rozlišeno, zda se jedná o právnickou nebo fyzickou osobu. Tato veličina je binární proměnnou, nabývá pouze dvou hodnot a to 0 nebo 1. Taktéž proměnná pohlaví je binární. Mezi kategorické veličiny v našem datovém souboru dále patří palivo, spoluúčast a kraj. Dalo by se říci, že i věk, cena nebo stáří vozidla by mohly být kategorickými veličinami, ale my pro zjednodušení předpokládáme, že jsou veličinami spojitými. Cena vozidel je velice diferencována, což můžeme říci podle výše směrodatné odchylky. Tento jev je způsoben vozidlem z námi aplikovaného datového souboru, které bylo koupeno za 8 mil. Kč.

Některé spojité veličiny lze zobrazit pomocí histogramu. Rozložení četnosti proměnné časová expozice je zobrazeno na Obr. 4.1.

Obr. 4.1 Histogram časové expozice

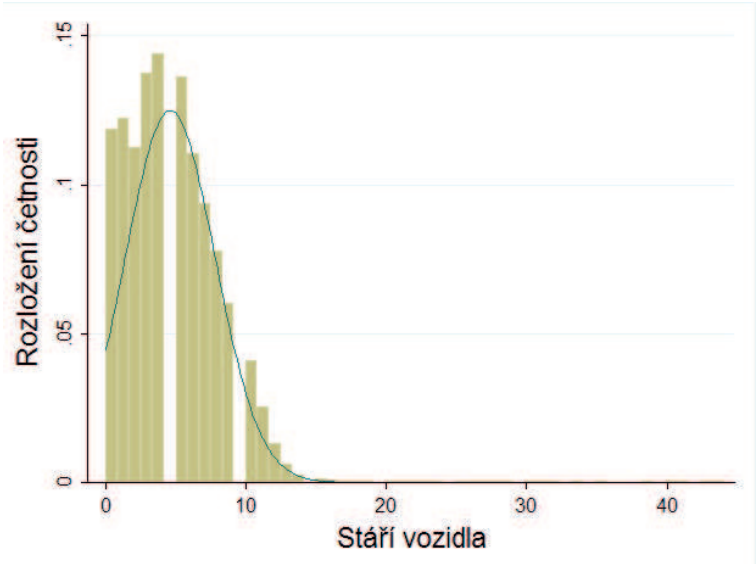


Zdroj: vlastní zpracování ve STATA 11.0

Z grafu je patrné, že nezávislá veličina nemá normální rozdělení pravděpodobnosti a je nejčetnější v bodě 1. To znamená, že většina pojistných smluv trvala celý rok.

V následujícím grafu je znázorněno rozložení četnosti u proměnné agecar- stáří vozidla, kde pro lepší srovnání jsme zobrazili, jak by měla vypadat proměnná s normálním rozdělením pravděpodobnosti. Histogram četnosti je na Obr. 4.2.

Obr. 4.2 Histogram stáří vozidla

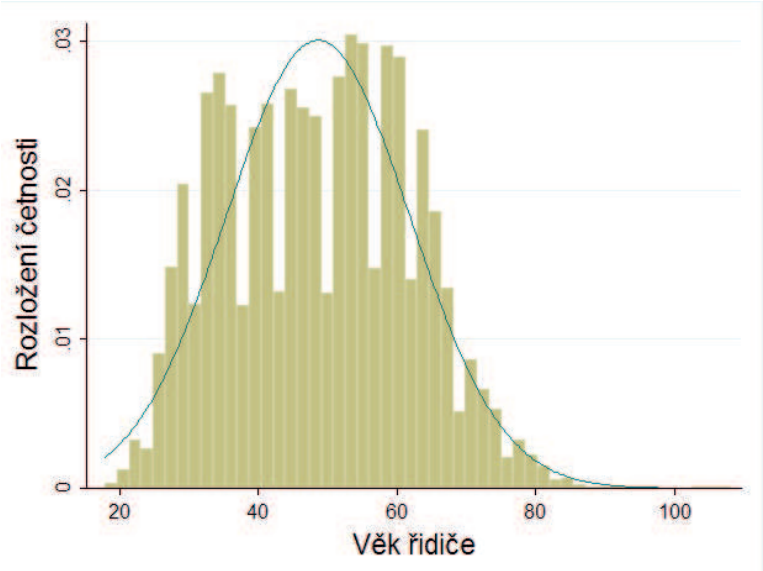


Zdroj: vlastní zpracování v STATA 11.0

Z grafu je patrné, že spojitá proměnná *stáří vozidla* nepochází z normálního rozdělení. Je špičatější, šikmější a nenabývá záporných hodnot. Nejčastěji dochází k nehodě, kdy vozidla jsou 5 let stará.

Druhou spojitou veličinou, u které je zobrazeno rozložení četnosti pomocí histogramu je proměnná ageman neboli věk řidiče. Rozložení četnosti je na Obr. 4.3.

Obr. 4.3 Histogram proměnné věk řidiče

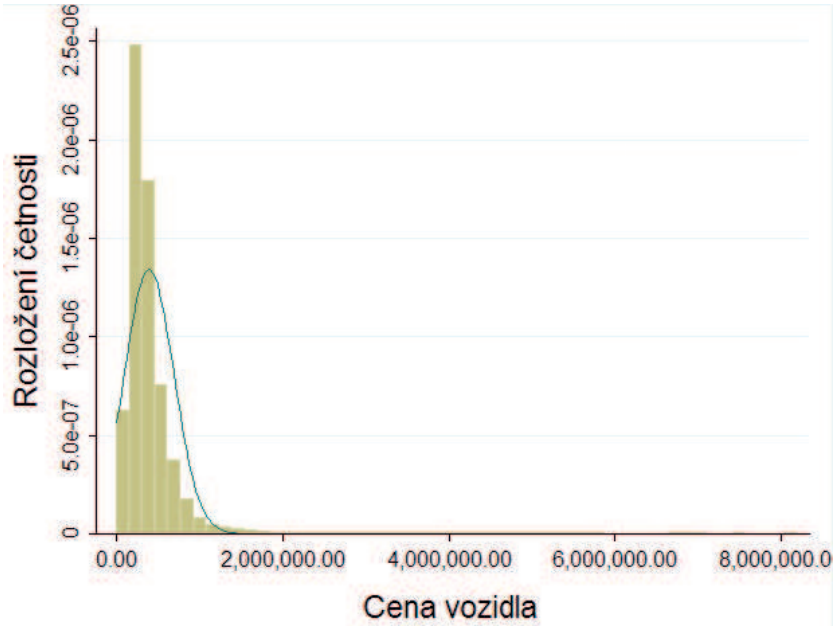


Zdroj: vlastní zpracování v STATA 11.0

U proměnné *věk řidiče* jsme rovněž zjistili, že se nejedná o normální rozdělení. Četnosti jsou různě rozložené a jsou mírně špičatější oproti normálnímu rozdělení.

Poslední rozložení četnosti je zobrazeno pomocí histogramu pro proměnnou price, čili cena. Touto proměnnou je představena cena, za kterou bylo koupeno pojištěné motorové vozidlo. Rozložení četnosti je na Obr. 4.4.

Obr. 4.4 Histogram proměnné cena vozidla



Zdroj: vlastní zpracování v STATA 11.0

Tato veličina taktéž nepochází z normálního rozdělení, nenabývá záporných hodnot a je mnohem špičatější než normální rozdělení. Z histogramu je patrné, že cena vozidla se nejčastěji pohybovala v intervalu od 100 tis. Kč do 500 tis. Kč.

4.2 Jednofaktorová analýza

Podkapitola je zaměřena na odhady parametrů negativně binomické regrese při pomoci užití metody maximální věrohodnosti. U jednotlivých odhadů je počítáno pouze s jedinou vysvětlující proměnnou a je zjištěno, zda vybraná proměnná je statisticky významná a měla by být použita k odhadu finálního modelu.

Nejdříve ovšem vybereme proměnné z datového souboru, se kterými budeme pracovat. Dle subjektivního uvážení se v diplomové práci nepoužívají proměnné, které souvisí s velikosti škody jako size, claim atd. Tyto proměnné dle našeho úsudku nemají vliv na četnost pojistných událostí. V souboru dat jsou zastoupeny rovněž proměnné jako průměrný věk v kraji, počet obyvatel v kraji, výše škody atd. Dle logického uvážení jsou tyto proměnné vynechány a nepředpokládáme, že četnosti pojistných událostí u havarijního pojištění jsou na těchto proměnných závislé. Po subjektivním vyfiltrování veličin ještě je nutné vyloučit vzájemnou kolinearitu mezi proměnnými. Je-li hodnota v absolutním vyjádření menší než 0,85, nejedná se o vzájemnou závislost mezi proměnnými. Multikolinearita je

zaznamenána mezi veličinami kw a volume. Tyto dva regresory jsou na sobě vysoce závislé, hodnota v absolutním vyjádření byla ve výši 0,9. Proto je vytvořena umělá proměnná kwvol jako součin kw a volume. Proměnná size neboli výše škody je taktéž vynechána z důvodu multikolinearity s proměnnou count, čili počet nehod. Všechny ostatní vysvětlující proměnné jsou podrobeny testování významnosti. Po vyloučení kolinearity mezi jednotlivými regresory, můžeme přejít k první fázi odhadování nejlepšího modelu a tou je jednofaktorová analýza. Cílem jednorozměrné analýzy je testování, zda proměnné jsou statisticky významné a mají vliv na vysvětlovanou proměnnou. Všechny veličiny jsou samostatně podrobeny analýze a následně posouzeny, zda mají být zahrnuty v modelu.

Prvním regresorem, u kterého je testovaná významnost pro odhad finálního modelu, je veličina company neboli typ ekonomického subjektu. Proměnná company je binární veličinou, nabývá hodnot nula nebo jedna. Proměnnou company zjistíme, zda havarijní pojištění je sepsáno na právnickou či fyzickou osobu. Pokud veličina subjekt je rovna nule, pojištěným je právnická osoba a pokud subjekt nabývá hodnoty jedna, jedná se o fyzickou osobu. Pro odhad koeficientu modelu je užíván software STATA 11.0. Do možnosti příkazu zapíšeme, že veličiny pochází z exponenciálního typu rozdělení a to negativně binomického a pro odhad je použita metoda maximální věrohodnosti. Pro odhad zobecněného lineárního modelu ve STATA 11.0 je zaveden Newton- Raphson algoritmus a jeho pozorovaná informační matice OIM. Je možné rovněž nadefinovat metodu IRLS s očekávanou informační matici EIM, ale my vycházíme ze základního nastavení. Všechny modely jsou odhadovány na hladině spolehlivosti 95% a je posuzovaná tak zvaná *z-statistika*. Při odhadech rovněž nesmíme zapomenout počítat s časovou expozicí, která má vliv na výsledný model. Výsledek odhadovaného modelu s proměnnou company je zobrazen v Tab. 4.2.

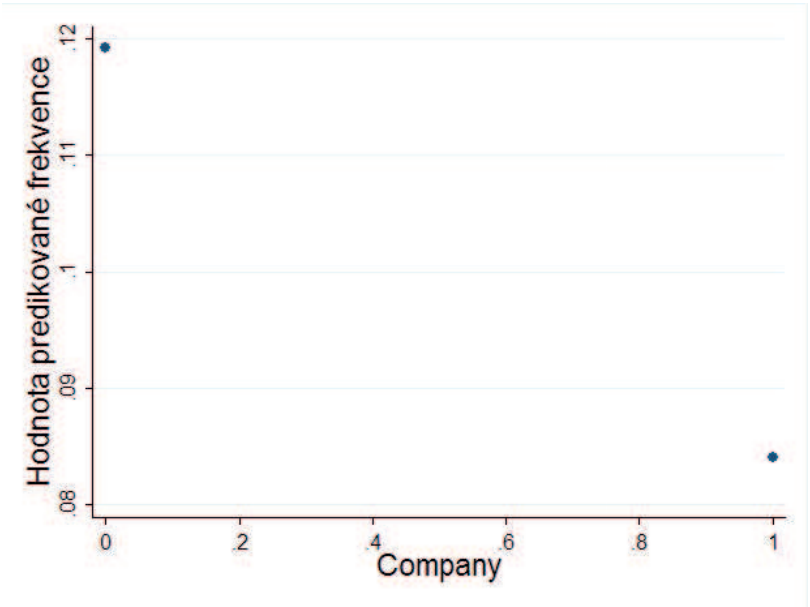
Tab. 4.2 Výsledky jednofaktorové analýzy jednotlivých proměnných

Proměnná	Koeficient	Směr. chyba odhadu	*z*-hodnota	Významnost	Dolní konf. interval	Horní konf. interval
Company	-0,350	0,017	-20,28	0,000	-0,383	-0,316
Gender	0,225	0,022	10,18	0,000	0,182	0,268
Diesel	0,580	0,018	32,77	0,000	0,545	0,615
P-butan	0,670	0,442	1,52	0,130	-0,197	1,537
Ostatní	0,327	0,294	1,11	0,267	0,250	0,903
Čas.expozice	-1,073	0,031	-34,88	0,000	-1,133	-1,013
Spol. 10%	-0,304	0,021	-13,85	0,000	-0,347	-0,261
Spol. 15%	-0,759	0,109	-6,96	0,000	-0,973	-0,546
Spol. 20%	-0,874	0,098	-8,87	0,001	-1,067	-0,681
Spol. 30%	-1,334	0,384	-3,47	0,000	-2,087	-0,581
Kwvol	0,025	0,001	26,77	0,000	0,023	0,027
Věk řidiče	-0,025	0,001	-24,88	0,000	-0,022	-0,019
Stáří vozidla	-0,048	0,003	-17,25	0,000	-0,053	-0,042

Zdroj: vlastní zpracování v STATA 11.0

Jednou z nejdůležitějších hodnot, která je výstupem z programu, je p- value neboli významnost, na základě které je rozhodováno, zda odhadovaný koeficient je statisticky významný. Zda zamítáme nulovou hypotézu nebo naopak. Koeficient je statisticky významný na hladině významnosti 5%, je-li p- value menší než 0,05. U proměnné company hodnotou z = -20,28 je vyjádřena pravděpodobnost nejméně extrémního výsledku a hodnota je umístěna v konfidenčním intervalu od - ∞ do -0,383. Směrodatná chyba odhadu nabývá malých hodnot, na hladině spolehlivosti 95%. Nízká je z toho důvodu, že máme relativně velký počet pozorování. Odhadovaný koeficient proměnné company je -0.349, má negativní vliv na vysvětlovanou proměnnou. Vliv koeficientu na četnost pojistných událostí je lépe znázorněn na Obr. 4.5.

Obr. 4.5 Vliv typu subjektu na frekvenci



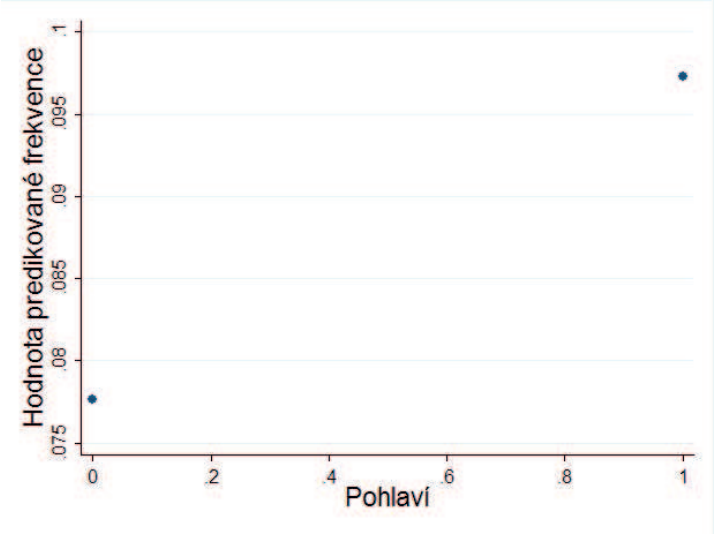
Zdroj: vlastní zpracování ve STATA11.0

V grafu je zobrazeno, že pokud pojištěným je právnická osoba, která je vyjádřena hodnotou nula, je větší pravděpodobnost vzniku pojistné události, než pokud pojištěným je fyzická osoba. Můžeme to vysvětlit tím, že pokud pojištěným je právnická osoba, čili nějaká firma, motorové vozidlo řídí zaměstnanci firmy a nestarají se tak o služební vozidla, jak o svá vlastní. Jízdou služebním vozidlem daleko více riskují. Prediktor company je statisticky významný a proto jej použijeme v odhadu finálního modelu.

Druhou proměnnou, jejíž koeficient je odhadován v jednofaktorové analýze, je veličina gender- pohlaví. Jedná se o binární proměnnou, kdy hodnotu jedna přiřadíme ženskému pohlaví a hodnotu nula mužskému. Hodnota nula je daleko častěji zastoupena v souboru dat, jelikož jako nula je započítána i právnická osoba. Proto při analýze s proměnnou pohlaví musíme rovněž do modelu zahrnout další proměnnou a to company. Bude-li company nula a gender nula, jedná se o právnickou osobu. Pokud hodnota proměnné company bude jedna a gender nula jedná se o fyzickou osobu mužského pohlaví. Odhadnutý model s veličinou pohlaví je zobrazen v Tab. 4.2.

Z tabulky jsme zjistili, že koeficient gender je opět statisticky významný a může být zahrnut do modelu. Pravděpodobnost je rovna nule, zamítáme nulovou hypotézu, že koeficient je roven nule. Pro lepší znázornění vlivu vysvětlující proměnné na vysvětlovanou proměnnou je Obr. 4.6.

Obr. 4.6 Vliv pohlaví na frekvenci nehod



Zdroje: vlastní zpracování v STATA 11.0

Při odhadu jsme počítáno pouze s fyzickými osobami a jsou vyeliminovány ze souboru dat údaje týkající se právnických osob. Z grafu je patrné, že větší počet nehod je způsoben řízením ženy. Zjištěním je popřen předpoklad, že ženy jezdí opatrněji a tím způsobují méně dopravních nehod. Odhadnutý koeficient je statisticky významný na pozorovaném vzorku dat a regresor gender bude zahrnut do finálního modelu.

Další proměnnou, u které budeme testovat, zda její vliv na četnost škod je významný, je fuel- palivo. Tato vysvětlující veličina je kategorickou proměnnou a je rozdělena do čtyř kategorií. Při odhadu pracujeme se čtyřmi typy paliv a to benzín, nafta, p-butan a poslední kategorií je položka ostatní viz Obr. 4.7.

Obr. 4.7 Typy paliv

Fuel category	Freq.	Percent	Cum.
petrol	266,709	73.63	73.63
diesel	95,250	26.30	99.93
p-butane	57	0.02	99.94
other	211	0.06	100.00
Total	362,227	100.00	

Zdroj: vlastní zpracování v STATA 11.0

Největší zastoupení má kategorie benzín (petrol) přes 73%, naopak nejméně je v souboru dat zastoupen propan butan. Jelikož se jedná o kategorickou veličinu, je potřeba

softwaru STATA dát příkaz, aby tento jev byl při odhadu zohledněn. Ze čtyř kategorií vzniknou tři a jako základ bude brána první kategorie benzín. Výsledek odhadu je v Tab. 4.2. Počet pozorování je stále 233 879. Z Tab. 4.2 je zřejmé, že dvě kategorie paliv jsou statisticky nevýznamné vzhledem k vybrané bázi. Při odhadech jsme poměnili různé báze a dvě kategorii byly stále nevýznamné. Před odstraněním těchto kategorii z modelu ještě je možné zjistit, zda veličina fuel- palivo je významná jako celek, k tomu slouží Waldův test. Nejdříve jsou stanoveny hypotézy dle vzorce (3.50) a (3.51). Tvrzení, že odhadnutý koeficient kategorické proměnné je nevýznamný, je určen v nulové hypotéze. Naopak alternativní hypotéza je založena na předpokladu, že koeficient je statisticky významný a je různý od nuly.

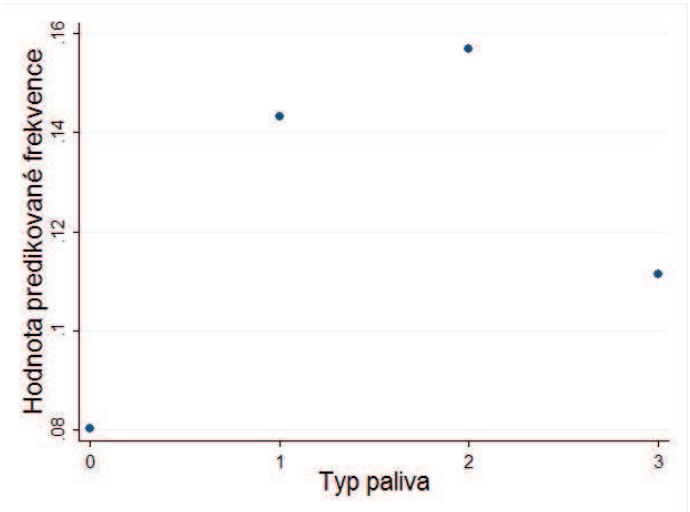
Tab. 4.3. Waldův test kategorické veličiny typ paliva

Chi^2 (3)	Významnost (Chi^2)
1075,32	0,000

Zdroj: vlastní zpracování v STATA 11.0

Pomocí Waldova testu jsme zjistili, zda je možné danou proměnnou v modelu vynechat. Vyjde-li pravděpodobnost menší než 0,05 na hladině spolehlivosti 95%, je zamítnuta nulová hypotéza, že koeficienty jsou rovny nule. Dle Tab. 4.3 regresor jako celek je statisticky významný na hladině významnosti 5% a proto je do finálního modelu zahrnut. Vliv jednotlivých kategorií na hodnotu predikované frekvence je zobrazen na Obr. 4.8.

Obr. 4.8 Vliv typu paliva na frekvenci nehod

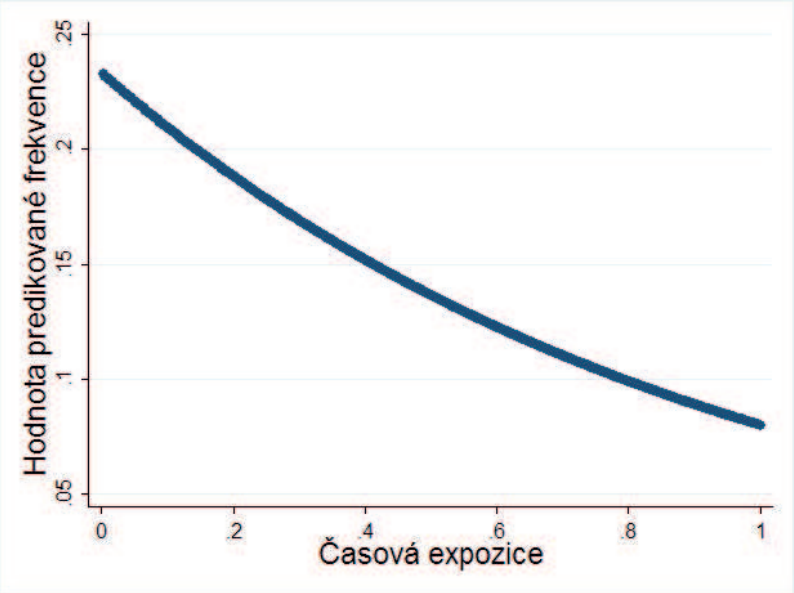


Zdroj: vlastní zpracování v STATA 11.0

Nejvyšší četnost pojistných událostí u havarijního pojištění je způsobena vozidly, která jezdí na plyn, dále pak na naftu a nejnižší škodovost jsme zjistili u benzínových motorových vozidel.

Jednou z dalších náhodných vysvětlujících proměnných je duration čili časová expozice. Touto veličinou je představena doba, jakou poměrnou část roku trvá pojištění. Výsledky jednofaktorové analýzy nezávislé proměnné jsou zobrazeny v Tab. 4.2. Proměnná duration taktéž je taktéž zahrnuta do celkového modelu, jelikož je statisticky významná na hladině spolehlivosti 95%. Z tabulky vidíme, že koeficient je záporný a pro lepší představu vliv proměnné duration je možné vidět na Obr. 4.9.

Obr. 4.9 Vliv doby trvání smlouvy na frekvenci pojistných událostí



Zdroj: vlastní zpracování v STATA 11.0

Na Obr. 4.9 je znázorněn vliv časové expozice na frekvenci nehod. Čím je delší doba trvání pojištění, tím je nižší četnost nehod. Což je logické, jelikož většinou když dojde k nějaké pojistné události, škoda je nahrazena a pojištění zaniká. Kdyby došlo v brzké době po uzavření havarijního pojištění k pojistné události na příklad do půl roku, ročně by to bylo dvakrát tolik.

Další významnou proměnnou, jež by mohla ovlivňovat frekvenci škod, je excess čili výše spoluúčasti na pojistném plnění. Spoluúčast patří mezi kategorické veličiny, její výše je buď 5%, 10%, 15%, 20% nebo 30% na škodě. Nejčastěji se vyskytuje pěti procentní spoluúčast. Problémem tohoto regresoru není pouze to, že se jedná o kategorickou veličinu,

ale že je rovněž ordinální. Bohužel jak to často v realitě bývá, je-li škoda v menší výši, pojištěnému se neoplatí hlásit pojistnou událost, aby nepřišel o výhodné bonusy a proto výsledek může být mírně zkreslený. Odhad koeficientů vybraných kategorií je v Tab. 4.4.

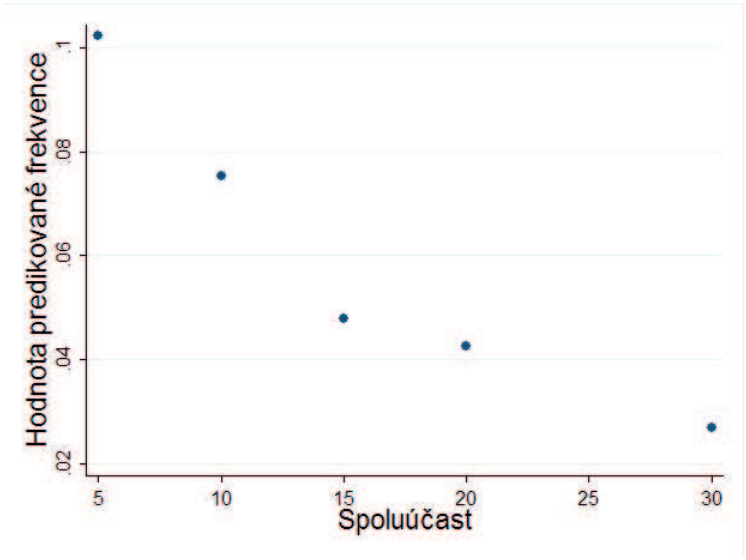
Tab. 4.4 Výsledky jednofaktorové analýzy u proměnné spoluúčast

Proměnná	Koeficient	Směr. chyba odhadu	*z*-hodnota	Významnost	Dolní konf. interval	Horní konf. interval
Spoluúčast 10%	-0,304	0,021	-13,85	0,000	-0,347	-0,261
Spoluúčast 15%	-0,759	0,109	-6,96	0,000	-0,973	-0,546
Spoluúčast 20%	-0,874	0,098	-8,87	0,001	-1,067	-0,681
Spoluúčast 30%	-1,334	0,384	-3,47	0,000	-2,087	-0,581

Zdroj: vlastní zpracování v STATA 11.0

V odhadnutém modelu všechny kategorie spoluúčasti na pojistném plnění jsou statisticky významné na hladině spolehlivosti 95%. Odhadované koeficienty jsou záporné, čím vyšší je spoluúčast na škodním průběhu, tím je frekvence nehod nižší viz Obr. 4.10.

Obr. 4.10 Vliv vysvětlující proměnné výše spoluúčasti na frekvenci nehod



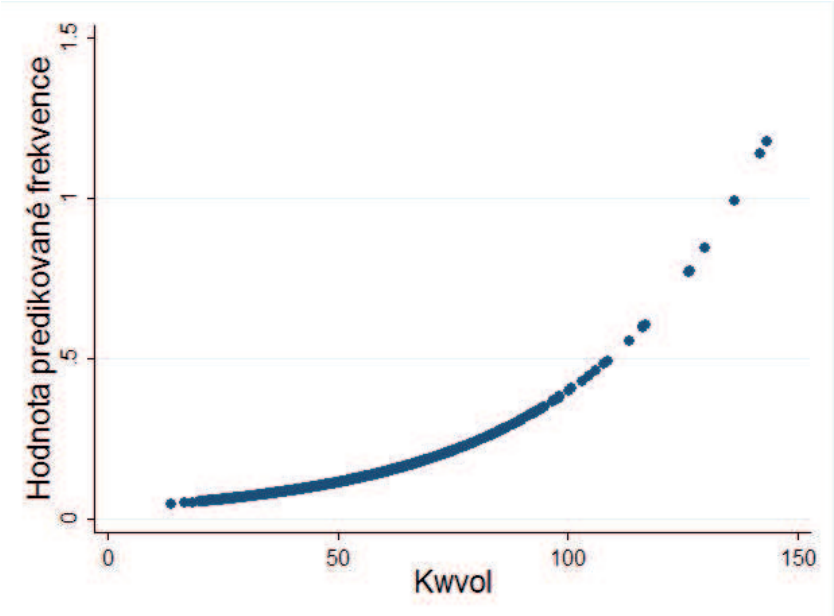
Zdroj: vlastní zpracování v STATA 11.0

Zde se ovšem naskytuje otázka, zda je to tím, že nehody s nižší škodou nejsou hlášeny pojišťovně a proto nejsou započítány v datovém souboru pojišťovny. Jelikož čím vyšší je spoluúčast a bezškodný průběh, tím větší bonusy na pojistném pojistníci mají a z toho důvodu menší škody raději nenahlásí. Druhá verze může být taková, že čím vyšší je spoluúčast, tím opatrněji pojištěny řídí své motorové vozidlo, aby nedošlo k pojistné události a nemusel se podílet na pojistném plnění. I když parametry jednotlivých kategorií jsou statisticky významné, raději veličina excess nebude zahrnuta do finálního modelu, protože by výsledky mohly být zkreslené. Výše spoluúčasti by měla vliv na kalkulaci pojistného a tam bychom tuto proměnnou museli do modelu zahrnout.

Na sílu motorového vozidla má vliv výkon a obsah, jelikož tyto dvě veličiny jsou multikolineární. Závislost mezi vysvětlujícími proměnnými je nad cca 0,9. Tyto dvě proměnné jsou na sobě vysoce závislé a z toho důvodu byla vytvořena „umělá“ proměnná kwvol jako: $kwvol = kw/volume \cdot 1000$. Budeme proto odhadovat model s vytvořenou proměnnou kwvol. Výstup ze STATA je v Tab. 4.2.

Proměnná kwvol dle *z-statistiky* je statisticky významná a může být použita ve finálním modelu. Mezi vysvětlovanou proměnnou a vysvětlující proměnnou je pozitivní závislost viz Obr. 4.11.

Obr. 4.11 Vliv výkonnosti vozidla na četnost nahodilostí



Zdroj: vlastní zpracování v STATA 11.0

Z Obr. 4.11 je patrná pozitivní závislost mezi proměnnými, čím vyšší je výkon a obsah motoru vozidla, tím častěji dochází k nehodám. Předpokládáme, že s výkonnějšími vozidly, řidiči více riskují, hazardují a u „silných“ automobilů je dosahováno většího zrychlení za kratší okamžik.

Jednou z dalších proměnných, u kterých je testována statistická významnost, je proměnná district. Jedná se o kraj, ve kterém se nachází sídlo právnické osoby nebo bydliště osoby fyzické. District je kategorickou veličinou a je rozdělena na čtrnáct kategorií. V našem státě máme čtrnáct krajů včetně Prahy. Kategorie proměnné district jsou zobrazeny v Tab. 4.5.

V následujícím odhadu Jihočeský kraj je vybrán jako základ.

Významnost jednotlivých kategorií krajů odhadovaného modelu je zobrazen v Tab. 4.5.

Tab. 4.5 Statistická významnost kategorií proměnné district

Kraj v ČR	Významnost
Královehradecký	0,000
Pardubický	0,000
Ústecký	0,617
Jihomoravský	0,00
Liberecký	0,104
Plzeňský	0,217
Zlínský	0,000
Karlovarský	0,090
Moravskoslezský	0,000
Praha	0,000
Vysočina	0,000
Olomoucký	0,003
Středočeský	0,499

Zdroj: vlastní výpočty v STATA 11.0

Po odhadu modelu jsme zjistili, že pět krajů je statisticky nevýznamných na hladině významnosti 5% a to Ústecký, Liberecký, Plzeňský, Karlovarský a Olomoucký kraj. Opětovně je významnost testována Waldovou statistikou, zda veličina district jako celek je statisticky významná a zda v modelu budou ponechány všechny kategorie vybrané veličiny. Nulovou hypotézou je stanoven předpoklad, že všechny odhadnuté parametry beta jsou rovny

nule dle vzorce (3.50), kdežto alternativní hypotéza je založena na předpokladu, že alespoň jeden koeficient je různý od nuly (3.51). Vyjde-li hodnota *významnosti (Chi^2)* menší než 0,05, je zamítnuta nulová hypotéza. Test je vypočten v programu STATA 11.0. Výsledky testu jsou v Tab. 4.6.

Tab. 4.6 Wald test proměnné district

Chi^2 (3)	Významnost (Chi^2)
548,35	0,000

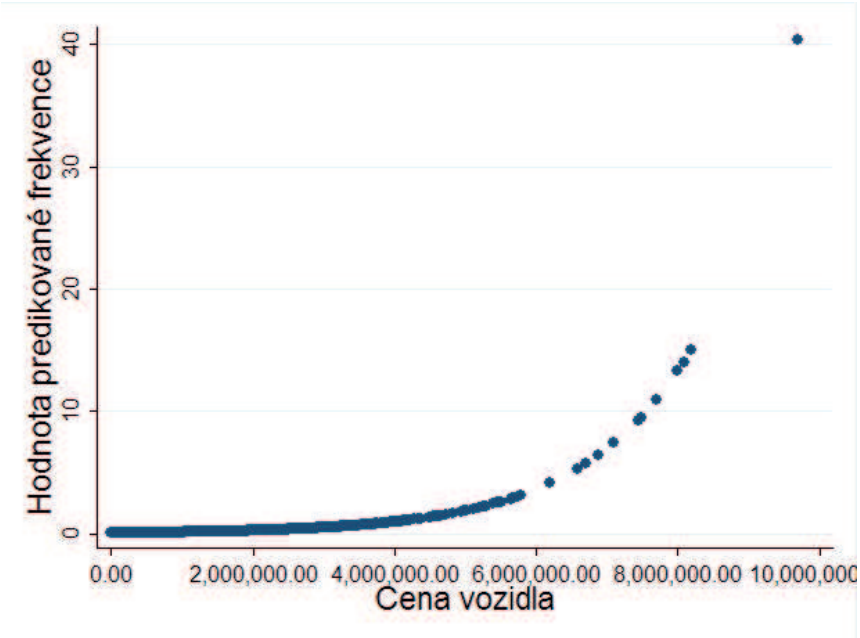
Zdroj: vlastní výpočet v STATA 11.0

Proměnná jako celek je významná a je důležitou součástí modelu, proto ji nebudeme vynechávat a zahrneme do finálního modelu všechny kategorie veličiny district (kraje).

Další proměnnou, kterou je podrobena jednofaktorové analýze, je price- cena motorového vozidla. Hodnota odhadnutého koeficientu je velmi nízká, taktéž i chyba odhadu.

Významnost odhadovaného koeficientu je rovna nule. Proměnná je statisticky významná a zahrneme ji do celkového odhadovaného modelu. Z odhadu je patrné, že mezi proměnnými je pozitivní vztah. Čím vyšší bude cena motorového vozidla, na nějž se pojištění vztahuje, tím častěji bude docházet k pojistným událostem. Pro lepší představivost, závislost zobrazíme na Obr. 4.12.

Obr. 4.12 Vliv ceny vozidla na predikovanou frekvenci nehod

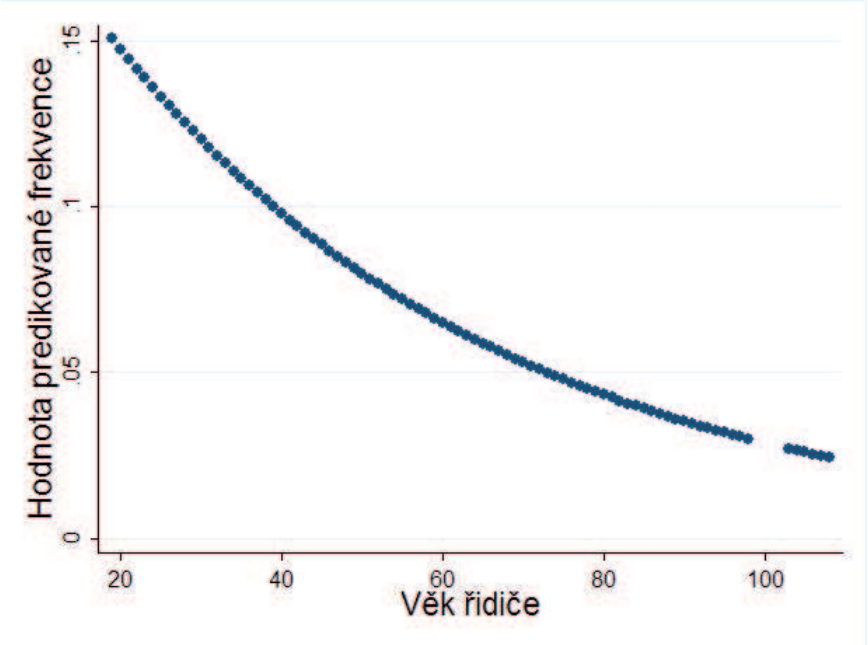


Zdroj: vlastní zpracování v STATA 11.0

V grafu je zřetelná závislost mezi vysvětlovanou a vysvětlující proměnnou. Když je cena vozidla menší než 4 mil. Kč, frekvence pojistných je téměř neměnná, ale když ceny vozidel už jsou vyšší, strmě roste pravděpodobnost nehody. Na Obr. 4.12 je zřejmý jeden osamocený bod. Cena havarovaného vozidla v tomto bodě je přes 9 mil Kč a během roku došlo u daného pojištěného k třem dopravním nehodám.

Dalším, dle mého názoru důležitým regresorem, kterým může být ovlivněna četnost pojistných událostí, je věk řidiče vozidla. Věkový rozsah je od 0 do 108 let. Musíme ale zdůraznit, že pod věkem 0 je bráno, že majitelem pojištěného vozidla není fyzická osoba, ale právnická. Z toho důvodu při odhadu závislosti frekvence nehod na věku řidiče budeme muset zohlednit i proměnnou company. Bude-li company 0 a ageman 0, jedná se o právnickou osobu a bude-li company 1, pojištěným je fyzická osoba. Výsledné hodnoty jsou uvedeny v Tab. 4.2. Závislost mezi vysvětlující proměnnou ageman a vysvětlovanou je negativní viz Obr. 4.13.

Obr. 4.13 Vliv věku řidiče na četnost pojistných událostí



Zdroj: vlastní zpracování v STATA 11.0

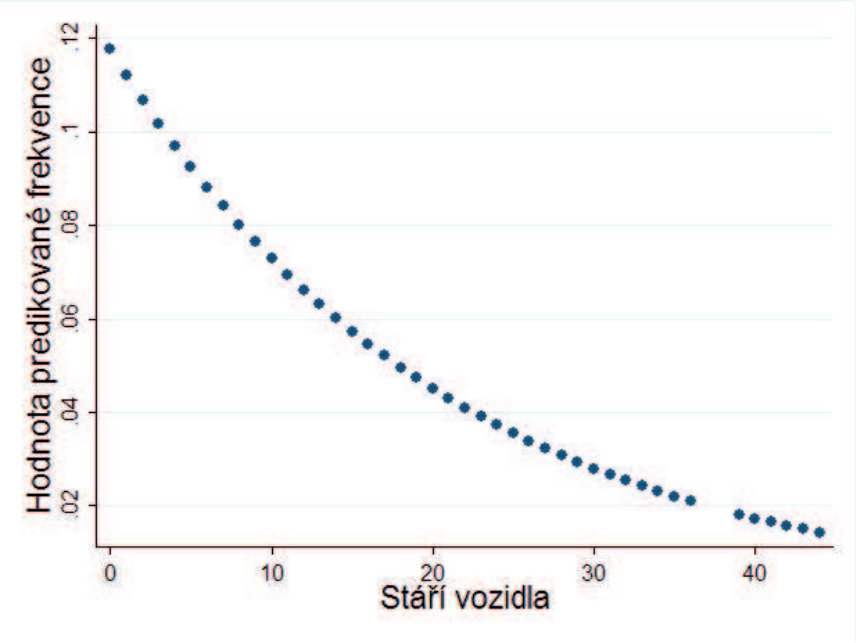
V grafu je jasně znázorněna negativní závislost. Čím vyššího věku dosahuje majitel pojištěného motorového vozidla, tím méně dochází k pojistným událostem. Teoretickými poznatky jsou potvrzovány poznatky z praxe. Většina mladých řidičů je nezkušená, ještě neví, jak správně zareagovat ve složitých situacích na cestách, ale rovněž mladí řidiči a to především mužského pohlaví více riskují. Mohli bychom rovněž namítat, že i řidiči

v pokročilejším věku tzn. nad 65 let, jsou nebezpečnější. Jelikož s vyšším věkem dochází ke ztrátě koncentrace, zhoršování zraku, zpomalení reflexů apod. Předpokládáme, že u starších řidičů, četnost nehod není taková vysoká, jelikož řidiči už tak často nejezdí a proto se i snižuje pravděpodobnost vzniku pojistné události.

Poslední proměnnou, u které je testována významnost parametrů, je stáří vozidla (agecar). Tento faktor bychom určitě neměli opomenout. Výsledné hodnoty odhadu zobecněného lineárního modelu pomocí metody MLE jsou uvedeny v Tab. 4.2.

Koeficient proměnné agecar je statisticky významný. Závislost pro lepší představivost je zobrazena na Obr. 4.14.

Obr. 4.14 Vliv stáří vozidla na hodnotu frekvence nehod



Zdroj: vlastní zpracování v STATA 11.0

Závislost je negativní, klesající, tedy s rostoucím stářím vozidla, klesá četnost nehod. Jednou z teorií je, že pokud dojde k nehodě u novějších vozidel, pojištěný nahlásí událost a škoda je uhrazena. U novějších vozidel je cena vyšší a proto i náhrada škody je vyšší. Druhou teorií je myšleno, že u starších vozidel pojištěný vzniklou pojistnou událost ani nenahlásí, jelikož náhrada škody by nebyla tak vysoká a on raději si škodu zaplatí z vlastních zdrojů, aby nepřišel o bonusy v bezeškodním průběhu. Další verzí je, že pokud dojde k pojistné události, je uhrazena škoda a pojištění u většiny případů zaniká. Proto logicky, čím je vozidlo starší a pojistná smlouva stále trvá, tím menší je frekvence pojistných událostí.

4.3 Odhady parametrů

Po odhadnutí koeficientů u jednofaktorových modelů bylo zjištěno, že všechny použité proměnné jsou statisticky významné a jsou zahrnuty celkového modelu. V dalším kroku je tedy nutné ověřit statistickou významnost odhadnutých koeficientů beta pomocí vícefaktorové analýzy. Finální model je taktéž odhadován metodou maximální věrohodnosti. Použitou optimalizační technikou je algoritmus Newton- Raphson, díky němuž je získána pozorovaná informační matice. Z diagonály negativně inverzní pozorované matice je zjištěna směrodatná chyba, která je potřebná pro určení významnosti modelu. Tím, že odhad je aplikován na velký počet pozorování, standardní chyby odhadu očekávané informační matice se příliš neliší od chyb pozorované matice. Pozorovaná matice OIM je vybrána z toho důvodu, že model NB2 je nekanonický, u nichž standardní chyby vypočítané z pozorované informační matice jsou méně zkreslené. Veličiny pocházejí z negativně binomického rozdělení, typ link funkce je logaritmický a model odhadujeme na hladině spolehlivosti 95%.

Nejdříve je odhadován model se všemi veličinami, které nám vyšly po jednofaktorové analýze významné. Následujícím kroku je model upraven o veličiny, které dle našeho názoru jsou v modelu zbytečné nebo jsou nevýznamné. Pomocí testu poměrem věrohodnosti LR dle vzorce (3.48) je zjištěno, který z vybraných odhadů je lepší pro modelování četnosti pojistných škod u havarijního pojištění. Rovněž pro určení vhodnějšího modelu je možné porovnat ukazatele AIC a BIC.

Finální odhadnutý model M1 se všemi významnými veličinami, které jsme odhadovali v jednorozměrné analýze, je uveden v Příloze č. 1. V Tab. 4.5 jsou zobrazeny pouze hodnoty, jež slouží ke srovnání mezi více modely. Mezi tyto patří informační kritéria nebo hodnota maximální věrohodnosti.

Pro odhad je použita již známa metoda maximální věrohodnosti a následně je odhadnutý model uložen do editoru jako M1, aby mohl být porovnán s jiným modelem testem věrohodnosti.

Hodnota zlogaritmovaného modelu věrohodnosti je -57 062. Hodnota Pearsnova rezidua je blízká jedné, což je žádaný jev, jelikož Pearsnovo reziduum vyděleno stupni volnosti by mělo být rovno jedné. Hodnoty informačních kritérií AIC a BIC jsou 0,4881 a -28178114. Některé z použitých veličin jsou na hladině významnosti 5% nevýznamné. Mezi nevýznamné veličiny patří tytéž kategorické proměnné, které vyšly nevýznamné i

v jednofaktorové analýze. Ale už v předchozích krocích použitím Waldovy statistiky bylo zjištěno, že jako celek, kategorie proměnné fuel a district významné jsou.

Finální model lze upravit (zredukovat) o proměnné fuel a district, které vyšly na hladině významnosti 5% statisticky nevýznamné. Je odhadován tzv. „neúplný“ model. Tento model rovněž uložíme pro budoucí LR test jako M2. Výsledky odhadu zobecněného lineárního neúplného modelu jsou uvedeny v Příloze č. 2. Po vyjmutí proměnných district a fuel je odhadnutý model statisticky významný. Hodnota AIC v tomto modelu je 0,489 a BIC je -2818585. V Tab. 4.7 jsou uvedeny pouze základní položky pro srovnání modelů M1 a M2.

Tab. 4.7 Srovnání odhadovaných modelů M1a M2

	Model M1	Model M2
Hodnota maximálně věrohodnostní funkce	-57 062,583	-57 171,855
AIC	114 173,2	114 367,7
BIC	114 421,9	114 492,1

Zdroj: vlastní zpracování v STATA 11.0

Srovnáme-li hodnoty informačních kritérií AIC a BIC vypočítaných dle vzorců (3.51) a (3.53) posledních odhadnutých modelů, zjistíme, že v „úplném“ finálním modelu jsou tyto hodnoty nižší. Dle teorie víme, že odhad s nižšími hodnotami informačních kritérií je vhodnější pro modelování četnosti. Použitím prvních srovnávacích kritérií vyšel „úplný“ model jako vhodnější model. Druhým testem, který slouží ke komparaci odhadnutých modelů, je test poměrem věrohodnosti dle vzorce (3.49). Tímto testem je srovnán složitější model s modelem zjednodušeným. LR testem lze zjistit, zda proměnné, které jsou vynechány v druhém odhadu, jsou významné a měly by být v modelu zařazeny. Nejdříve musí být stanoveny hypotézy. Nulová hypotéza je stanovena dle vzorce (3.50), že odhadované koeficienty proměnných, které jsme chtěli v modelu vynechat, jsou rovny nule, čili nevýznamné. Naopak alternativní hypotézou je myšleno, že všechny koeficienty jsou statisticky významné, čili jsou nenulové, viz vzorec (3.51). Důležitou hodnotou je *p-value* > chi^2, jež vypovídá o statistické významnosti vynechaných veličin. Je-li tato hodnota menší než 0,05, „úplný“ model je významný a měli bychom ponechat všechny vysvětlující proměnné a dále neaplikovat redukovaný model. Pro tento test jsme rovněž použili software STATA 11.0. Výsledky testu jsou zobrazeny v Tab. 4.8.

Tab. 4.8 Test poměrem věrohodnosti

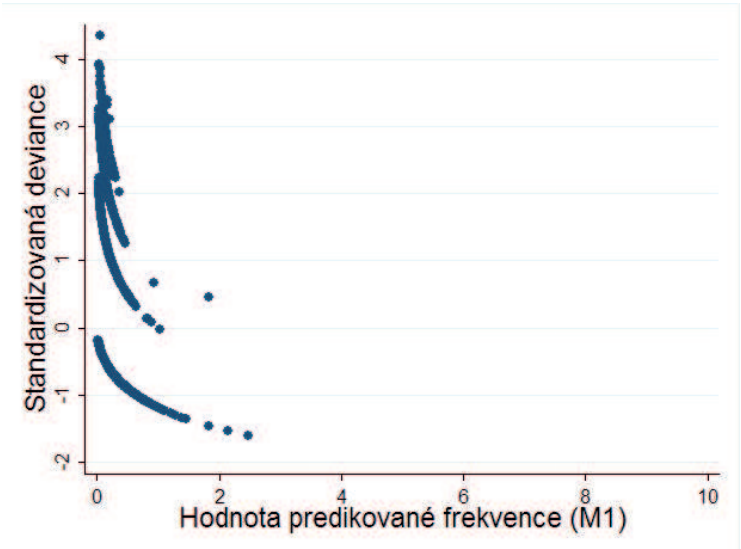
LR chi^2 (12)	Významnost > chi^2
218,54	0,000

Zdroj: vlastní zpracování v STATA 11

Model M2 je „umístěn“ do modelu M1, to znamená, že druhý model je zredukovaný model M1. Významnost vysvětlujících proměnných je testována LR testem (test poměrem věrohodnosti). Hodnota LR χ^2 rozdělení s 12 stupni volnosti je přes 218, to znamená, že veličiny, které jsme chtěli v modelu M2 vynechat, mají velký vliv na vysvětlovanou proměnnou. Hodnota pravděpodobnosti je menší než 0,05 na hladině významnosti 5%. Je zamítnuta nulová hypotéza, že odhadované koeficienty vysvětlujících proměnných jsou rovny nule. Koeficienty jsou tedy statisticky významné a měly by být k dohadu v modelu ponechány.

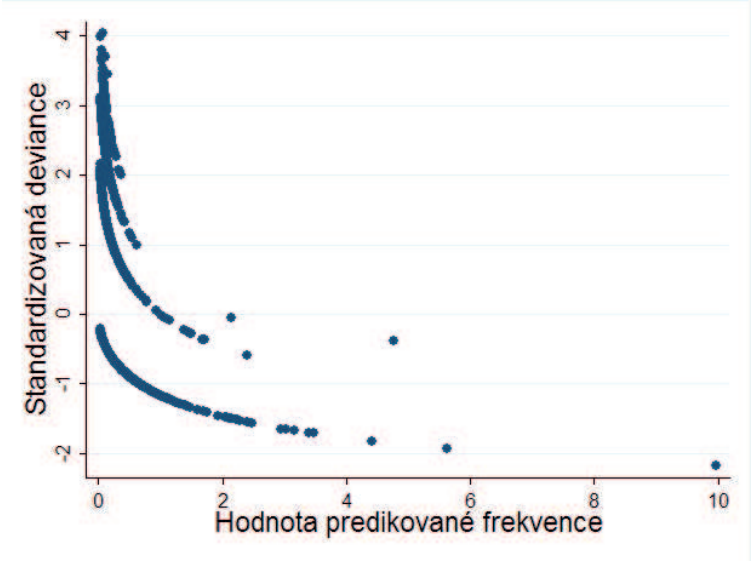
Modely M1 a M2 je možné rovněž srovnat pomocí grafického znázornění. U každého z modelů nejdříve po odhadu predikujeme standardizovaná rezidua deviance a predikovanou střední hodnotu četnosti pojistných událostí. Následně tyto dvě veličiny zobrazíme graficky a oba grafy porovnáme. Standardizovaná deviance odhadovaného modelu M1 jsou zobrazena graficky na Obr. 4.15 a standardizovaná deviance odhadovaného modelu M2 jsou na Obr. 4.16.

Obr. 4.15 Predikovaná rezidua deviance M1



Zdroj: vlastní zpracování

Obr. 4.16 Predikovaná rezidua deviance M2



Zdroj: vlastní zpracování

Z grafů je patrné, že v modelu M1 se vyskytuje více reziduí v intervalu od -2 do 2, což je ve srovnání s modelem M2 lepší. Hodnota odhadované frekvence v prvním modelu je daleko nižší než v modelu druhém. V modelu M2 je rovněž vyznačena extrémní hodnota frekvence nehod. Po grafickém srovnání obou modelů je lepší první odhadovaný model M1.

Po srovnání modelů M1 a M2 testem poměru věrohodnosti a srovnáním informačních kritérií jsme došli k závěru, že kategorické veličiny district a fuel ve finálním modelu budou ponechány. Rovněž v dřívější podkapitole pomocí Waldova testu bylo zjištěno, že proměnné jako celek jsou statisticky významné a jejich vynechání by mělo podstatný vliv na výsledky odhadovaného modelu. Kdyby finální model nebyl odhadován s těmito regresory, výsledky by byly zkreslené a neúplné.

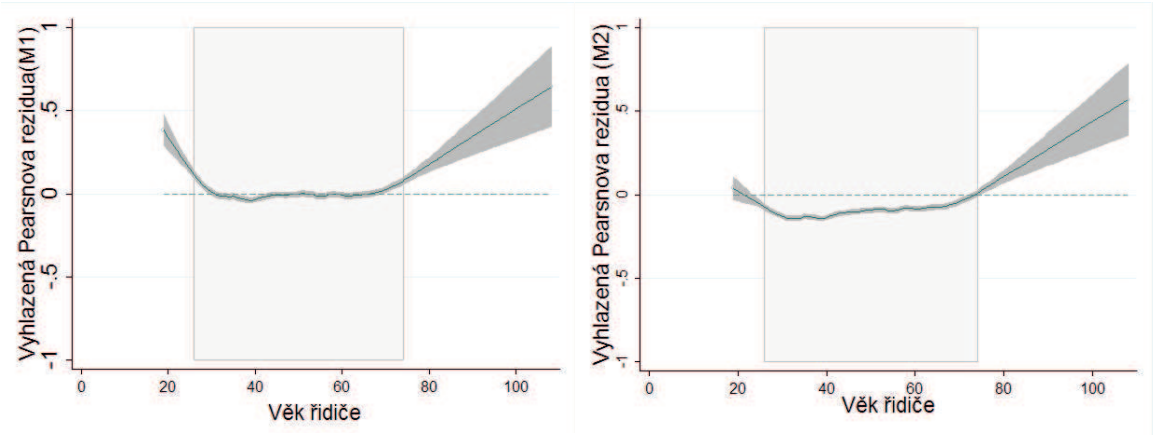
Modely je možno srovnávat taktéž pomocí *vyhlazených Pearsnových reziduí*. Rezidui je představena forma variability modelu. Jelikož rezidui je myšlena chybu odhadu, cílem je, aby tato rezidua byla jak nejmenší. Všechna vyhlazená rezidua by měla v ideálním případě být rovna nebo velmi blízká nule. Jestli je nějaká hodnota různá od nuly, model není ideální. To znamená, že čím víc jsou rezidua absolutně větší, tím víc se predikované hodnoty liší od reálných. Čím menší jsou odchylky od horizontální osy v bodě nula, tím je lepší odhadovaný model.

Nejdříve je nutno predikovat Pearsnova rezidua pro model M1 a M2 a následně tyto predikované hodnoty vyhladit jednotlivými spojitými proměnnými aplikovanými

v odhadovaném modelu. Vyhlazováním reziduí rovněž určíme hodnoty intervalů spolehlivosti jak dolních, tak i horních pro jednotlivé bodové odhady. Pearsnova rezidua jsou vyhlazována vždy oproti nějaké vysvětlující proměnné. V obou modelech je vyhlazováno pomocí spojitých proměnných jako věk řidiče, cena vozidla, časová expozice a stáří vozidla. Intervaly spolehlivosti jsou u obou modelů odhadovány na hladině spolehlivosti 95%. V následujících grafech jsou porovnány jednotlivé bodové odhady reziduí. Na ose y jsou vyznačeny hodnoty vyhlazených reziduí, škála na ose y je v rozmezí od -1 do 1. Na níže zobrazených grafech jsou predikovaná Pearsnova reziuda vyhlazená všemi spojitými proměnnými z modelu. Šedým obdélníkem v grafech je zvýrazněná oblast, ve které by se mělo nacházet 95% pozorování ve věkové kategorii od 18 do 75 let, to znamená, že mimo tuto oblast se nachází extrémní hodnoty datového souboru. V následujících grafech jsou zobrazeny hodnoty predikovaných Pearsnových reziduí vyhlazených podle dalších spojitých proměnných, se kterými je model odhadován.

Na Obr. 4.17 jsou zobrazena vyhlazena Pearsonova rezidua vůči proměnné věk řidiče z odhadovaných modelů M1 a M2.

Obr. 4.17 Srovnání vyhlazených reziduí proměnnou věk řidiče modelů M1 a M2



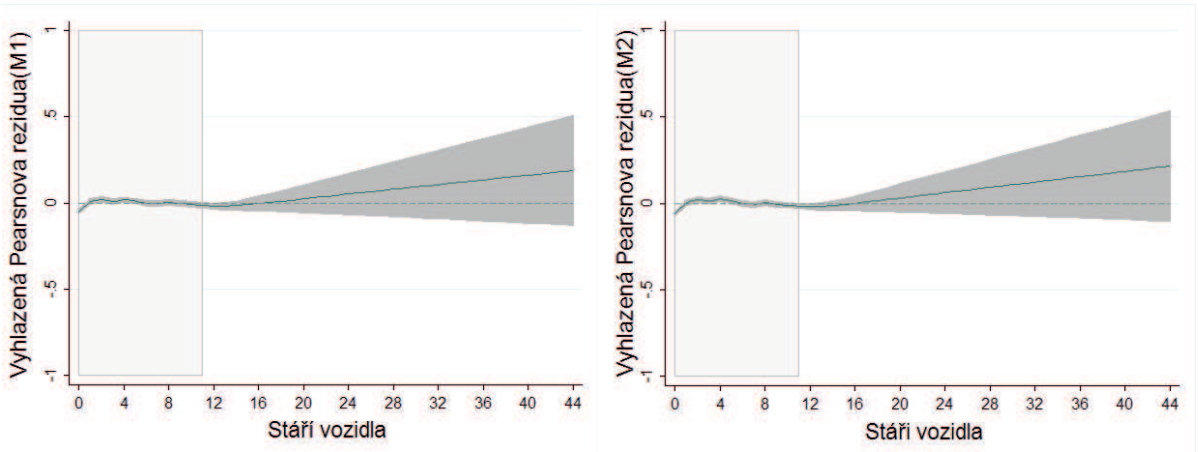
Zdroj: vlastní zpracování

Na první pohled je patrné, že očekávané hodnoty vyhlazených Pearsnových reziduí modelu M1 jsou pro modelování četnosti vhodnější než u modelu M2. Očekávané hodnoty reziduí korespondujících k lineární funkci vyznačené v šedém obdélníku (95% všech pozorování) jsou významné blíže nule. Pro lepší porovnání, jak moc se rezidua odchylují od správné hodnoty, je v grafu položena horizontální osa v bodě nula. V tomto grafu je rovněž vymezen interval, kde by se mělo nacházet 95% pozorování, šedým obdélníkem. Z Obr. 3.17 je patrné, že predikovaná a následně vyhlazena Pearsnova rezidua modelu M2 jsou od

horizontální osy mnohem víc vychýlená a minimum hodno je rozloženo kolem nulové osy. Rovněž extrémní hodnoty modelu jsou daleko více vzdálená než v modelu M1.

V následujícím Obr. 4.18 jsou zobrazena vyhlazená rezidua proměnnou stáří vozidla (agecar).

Obr. 4.18 Srovnání vyhlazených reziduí proměnnou stáří vozidla modelů M1 a M2

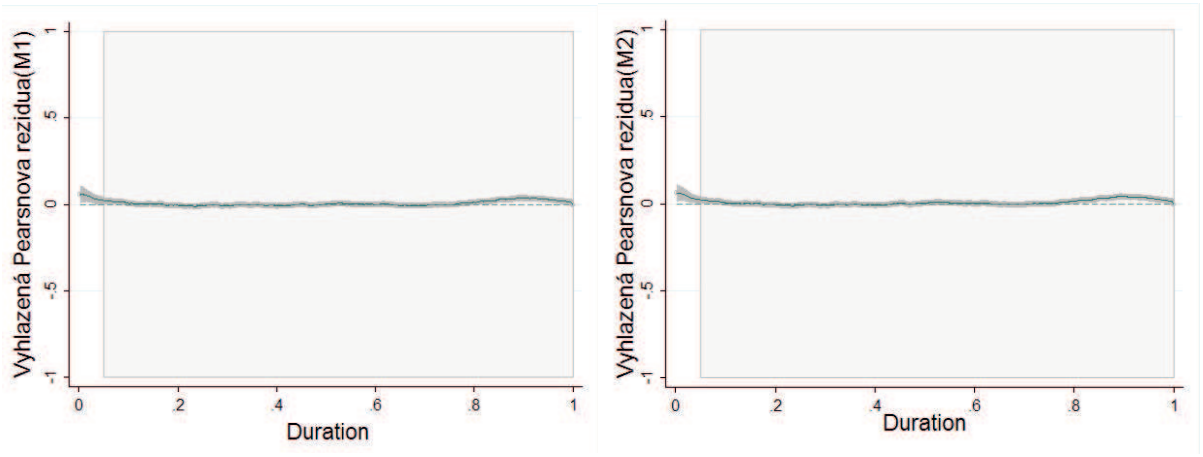


Zdroj: vlastní zpracování

Vyhlazená rezidua modelů M1 a M2 pomocí proměnné stáří vozidla se nějak významně neliší. Rovněž bychom mohli konstatovat, že vyhlazování proměnnou stáří vozidla je statisticky významnější- očekávané hodnoty reziduí jsou blíže nule, než vyhlazování proměnnou věk řidiče, kde hodnoty reziduí u extrémních pozorování byly výrazně vzdálené od nuly.

Na Obr. 4.19 jsou srovnána vyhlazená Pearsnova rezidua proměnnou duration (časová expozice) u obou odhadovaných modelů.

Obr. 4.19 Srovnání vyhlazených reziduí proměnnou duration modelů M1 a M2

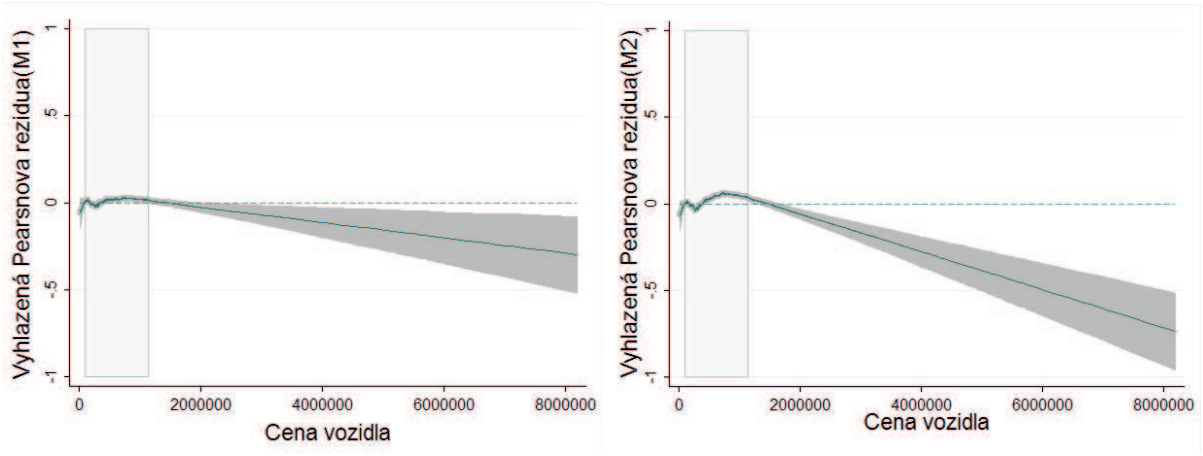


Zdroj: vlastní zpracování

Jak v předchozím případě očekávané hodnoty reziduí jak v modelu M1, tak i v modelu M2 jsou velmi podobné a rovněž statisticky významné- predikované hodnoty Pearsnových reziduí včetně hodnot v 95 % intervalech spolehlivostí. Rezidua vyhlazená proměnnou dauration jsou vhodná pro určení vhodnosti odhadovaného modelu pro modelování četnosti.

Na dalším Obr. 4.20 jsou vyznačeny predikované hodnoty vyhlazených reziduí poslední vysvětlující proměnnou cena vozidla.

Obr. 4.20 Srovnání vyhlazených reziduí proměnnou cena vozidla modelů M1 a M2



Zdroj: vlastní zpracování v STATA 11.0

Z Obr. 4.20 je patrné, že pro modelování četnosti je vhodnější model M1. Predikovaná Pearsnova rezidua se významně blíží hodnotě nula, čili jsou statisticky významnější než v modelku M2. Ale očekávané hodnoty reziduí u extrémních pozorování jsou v obou případech hodně vzdálené od nuly.

Po srovnání všech grafů odhadovaných modelů M1 a M2, jsme došli k závěru, že odhad M1 je daleko vhodnější pro modelování četnosti pojistných událostí havarijního pojištění. Rovněž tímto grafickým srovnáním se nám potvrdily výsledky Waldova testu a testu poměrem věrohodnosti, kde jsme zkoumali významnost koeficientů vybraných veličin. Pro modelování četnosti nahodilostí u havarijního pojištění je vhodnější model zahrnující proměnné kraj a palivo.

4.4 Kategorizace veličin

V další podkapitole je srovnán finální odhadovaný model M1 s novým odhadovaným modelem, kde vysvětlujícími proměnnými jsou pouze kategorické veličiny.

Prvním krokem je překódování všech spojitých veličin na veličiny kategorické. Binární proměnné jako company nebo gender nemusí být kategorizovány, jelikož binární veličina je brána jako specifický druh kategorické proměnné. Vliv na vysvětlovanou proměnnou je zkoumán pouze těmi vysvětlujícími proměnnými, které jsme již použili v modelu M1. Kategorizovány jsou pouze ty proměnné, kterých parametry byly dohadovány v modelu M1, aby bylo možné objektivně porovnat vliv kategorizovaných proměnných na četnost pojistných událostí. Nejdříve je kategorizována postupně každá spojitá proměnná zvlášť a následně pro statistickou významnost kategorizované proměnné je vypočtena jednofaktorová analýza. Je důležité ale poznamenat, že kategorizováním proměnných může dojít ke ztrátě některých podstatných informací.

První kategorizovanou veličinou je věk řidiče, čili ageman. Jako první je nutno zjistit základní poznatky o proměnné a to počet pozorování, střední hodnotu, směrodatnou odchylku, minimum a maximum z Tab. 4.1. Druhým krokem je rozdělení spojité veličiny do několika kategorií. Existuje více způsobů rozdělení spojitých veličin do kategorie, ale v tomto případě jsme vybrali tento způsob, kdy si sami určíme, jaké má být rozpětí jednotlivých kategorií. Dle subjektivního uvážení byly vytvořeny čtyři kategorie, první kategorie je tvořena řidiči ve věku od 18 do 25 let, druhou kategorii tvoří řidiči starší 25, ale mladší 50 let. Ve třetí kategorii jsou všichni řidiči v intervalu od 50 let včetně do 65 let a v poslední kategorii jsou osoby starší 65 let. Tento typ kategorizace jsme použili z toho důvodu, že jsme chtěli určit hranice subjektivně. Většina pozorování se nachází ve věkové kategorii od 20 do 65 let a procentuální zastoupení hodnot mimo tento interval je nízké. Četnost mezi jednotlivými kategoriemi je rozložena rovnoměrně. Dále vysvětlující proměnnou podrobíme

jednofaktorové analýze. Odhad modelu s jednou kategorizovanou proměnnou ageman je v Tab. 4.9.

Tab. 4.9 Odhady parametrů kategorizované proměnné věk řidiče

Kategorie	Koeficient	Směr. chyba odhadu	*z*-hodnota	Významnost	Dolní konf. Interval	Horní konf. Interval
<25;50)	-0,947	0,055	-17,15	0,000	-1,055	-0,839
<50;65)	-1,326	0,056	-23,35	0,000	-1,437	-1,215
<65;108)	-1,365	0,065	-21,03	0,000	-1,493	-1,238

Zdroj: vlastní zpracování v STATA 11.0

Na základě výsledků z Tab. 4.9 vyšlo, že všechny kategorie jsou na hladině významnosti 5% statisticky významné, jako základ je brána první kategorie, kdy všichni řidiči jsou mladší než 25 let.

Druhá proměnná, která je kategorizována, je uměle vytvořena proměnná kwvol. Nejdříve opět zjistíme minimální a maximální hodnotu. Minimum je 13,8 a maximum 143,3. Rovněž je použito subjektivní roztřídění proměnné kwvol. První kategorií jsou vozidla s výkonem do 35, další kategorie je tvořena intervalem <35;40), ve třetí kategorii jsou motorová vozidla s výkonem od 40 do 50, následně od 50 do 60 a do poslední kategorie jsou zařazena nejsilnější vozidla s výkonem nad 60. Po kategorizaci proměnné je testována významnost regresoru. Odhad s vytvořenou kategorickou proměnnou kwvol je v Tab. 4.10. Základ je tvořen opět první kategorii hodnot kwvol menších než 35.

Tab. 4.10 Odhady parametrů kategorizované proměnné kwvol

Kategorie	Koeficient	Směr. chyba odhadu	*z*-hodnota	Významnost	Dolní konf. interval	Horní konf. interval
<35;40)	-0,198	0,027	-7,43	0,000	-0,251	-0,146
<40;50)	0,278	0,025	11,27	0,000	0,229	0,326
<50;60)	0,493	0,026	18,72	0,000	0,441	0,544
<60;143,3)	0,795	0,070	11,29	0,000	0,657	0,933

Zdroj: vlastní zpracování v STATA 11.0

Všechny vytvořené kategorie veličiny kwvol vyšly statisticky významné a je možné všechny třídy zahrnout do celkového odhadu.

Další spojitá veličina, která je rozdělena do několika kategorií, je proměnná agecar neboli stáří motorového vozidla. Z tabulky lze zjistit minimální a maximální hodnotu vybrané veličiny. Minimální hodnota je 0 a maximální stáří vozidla je 44 let. Následně je veličinu kategorizována, kde první kategorii tvoří vozidla mladší než dva roky, druhá kategorie je tvořena vozidly, jejichž stáří je od 2 do 4 let. Ve třetí kategorii jsou zahrnutá vozidla v intervalu <4;7). Další kategorie je tvořena vozidly starší sedm let včetně, ale mladší než deset. V poslední kategorií jsou umístěna všechna vozidla starší více jak deset let. Toto rozdělení je rovněž subjektivní, ale rozložení četnosti v jednotlivých kategoriích je relativně rovnoměrné. Odhad modelu četnosti pojistných událostí při použití jedné kategorizované proměnné agecar je v Tab. 4.11.

Tab. 4.11 Odhad parametrů kategorizované veličiny stáří vozidla

Kategorie	Koeficient	Směr. chyba odhadu	*z*-hodnota	Významnost	Dolní konf. Interval	Horní konf. interval
<2;4)	0,015	0,022	0,66	0,510	-0,029	0,058
<4;7)	-0,175	0,022	-8,03	0,000	-0,217	-0,132
<7;10)	-0,342	0,028	-12,28	0,000	-0,397	-0,288
<10;44)	-0,605	0,053	-11,49	0,000	-0,708	-0,501

Zdroj: vlastní zpracování v STATA 11.0

Jako báze je dle nastavení taktéž v tomto případě vybrána první kategorie s vozidly mladšími než dva roky. Až na jednu kategorii jsou všechny kategorie statisticky významné na hladině spolehlivosti 95%. Tato hladina spolehlivosti je používána ve všech odhadech. První kategorie by byla významná na hladině významnosti větší než 51%, co už není žádané. Ale jako celková proměnná je agecarcat statisticky významná, což je potvrzeno Waldovým testem. Hodnota významnosti chi^2 rozdělení je 0,000 a je přijata alternativní hypotéza, že alespoň jeden odhadovaný parametr je statisticky významný.

Jednou ze spojitých proměnných je i cena- price. Rovněž tato veličina musí být kategorizována a následně je odhadován její parametr v jednorozměrném modelu. V softwaru STATA jsou vygenerovány jednotlivé kategorie. Je požadováno rozdělení do šesti tříd. Nejvíce vozidel v námi používaném datovém souboru je levnějších než 500 tis. Kč. V této cenové relaci je evidováno přes 77% motorových vozidel havarijního pojištění. Výsledky odhadu jednofaktorového modelu jsou v Tab. 4.12.

Tab. 4.12 Odhad parametrů kategorizované veličiny cena

Kategorie	Koeficient	Směr. chyba odhadu	*z*-hodnota	Významnost	Dolní konf. interval	Horní konf. Interval
<150000;230000)	0,006	0,042	0,15	0,882	-0,076	0,088
<230000;310000)	-0,169	0,037	-4,50	0,000	-0,242	-0,095
<310000;390000)	0,047	0,038	1,22	0,224	-0,028	0,122
<390000;470000)	0,252	0,041	6,10	0,000	0,171	0,332
<470000;55000)	0,557	0,036	15,61	0,000	0,487	0,627

Zdroj: vlastní zpracování v STATA 11.0

Základem pro tento odhad s kategorickou veličinou price je první kategorie s vozidly, jejichž cenová hladina je nižší než 150 tis. Kč. Dva odhadované koeficienty jednotlivých kategorií jsou statisticky nevýznamné na hladině spolehlivosti 95%. Z toho důvodu je potřebné významnost modelu jako celku otestovat Waldovou statistikou dle vzorce (3.52). V nulové hypotéze předpokládáme, že všechny odhadované beta koeficienty jsou nevýznamné (nulové) dle vzorce (3.50), kdežto alternativní hypotéza je založena na předpokladu viz vzorec (3.51), že alespoň jeden odhadovaný parametr je nenulový-významný. Po aplikaci testu je zjištěno, že kategorická proměnná jako celek je statisticky významná, hladina významnosti je 0,000, všechny koeficienty jsou tedy statisticky významné.

Poslední spojitou proměnnou, která je rozdělená do několika kategorií, je veličina duration neboli časová expozice. Minimum a maximum hodnot duration v datovém souboru je 0,0027 a 1. Maximum je logické, jelikož za rok může pojistná smlouva trvat maximálně rok. Pro kategorizaci proměnné duration jsme veličinu rozdělili do pěti kategorií. Něco přes 45% pojistných smluv během roku trvá víc než 9 měsíců. Odhad modelu je zobrazen v Tab. 4.13.

Tab. 4.13 Odhad parametrů kategorizované veličiny časová expozice

Kategorie	Koeficient	Směr. chyba odhadu	*z*-hodnota	Významnost	Dolní konf. interval	Horní konf. Interval
<0,208;0,406)	-0,213	0,051	-4,16	0,000	-0,313	-0,113
<0,406;0,604)	-0,390	0,049	-7,97	0,000	-0,486	-0,294
<0,604;0,802	-0,589	0,049	-12,02	0,000	-0,685	-0,493
<0,802;1)	-0,903	0,044	-20,72	0,000	-0,988	-0,817

Zdroj: vlastní zpracování v STATA 11.0

Všechny kategorie vysvětlující proměnné duration jsou statisticky významné a mohou být použity k odhadu modelu s kategorickými nezávislými veličinami.

4.5 Odhad modelu s kategorickými proměnnými

V této části je opět modelována četnost pojistných událostí havarijního pojištění. Jako regresory jsou v odhadu použity pouze kategorické veličiny jako agecat, agecarcat, fuel, durcat, kwvolcat, district a pricecat. Rovněž je model odhadován i s binárními proměnnými subjekt (company) a pohlaví (gender). Zobecněny lineární model je odhadován metodou maximální věrohodnosti. Odhad modelu je uveden v Příloze č. 3.

Po odhadu modelu jsme zjistili, že některé kategorie jednotlivých veličin jsou na hladině spolehlivosti 95% statisticky nevýznamné. Tyto nevýznamné kategorie nebudeme z modelu vynechávat, jelikož u jednofaktorových analýz byla u kategorických proměnných testována významnost jednotlivých veličin jako celku a všechny tyto veličiny vyšly na hladině významnosti 5% významné. Výše odhadnutý model je pro následující srovnání s modelem *M1* uložen jako *Mi*. Modely jsou porovnány testem poměru věrohodnosti dle vzorce (3.49), výsledek testu je uveden v Tab. 4.14.

Tab. 4.14 LR test modelů M1 a Mi

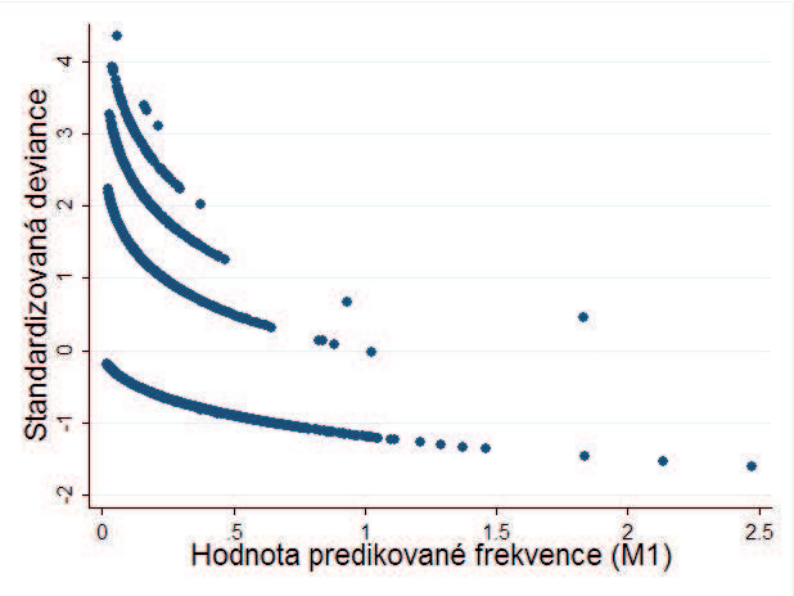
LR chi^2 (12)	Významnost> chi^2
68,9	0,000

Zdroj: vlastní výpočty v STATA 11.0

Po testování bylo zjištěno, že model s kategorizovanými veličinami jako celek je statisticky významný. Spojité veličiny, které byly kategorizovány, jsou při odhadu celkového modelu významné na hladině spolehlivosti 95%. I když hodnota ukazatele poměru

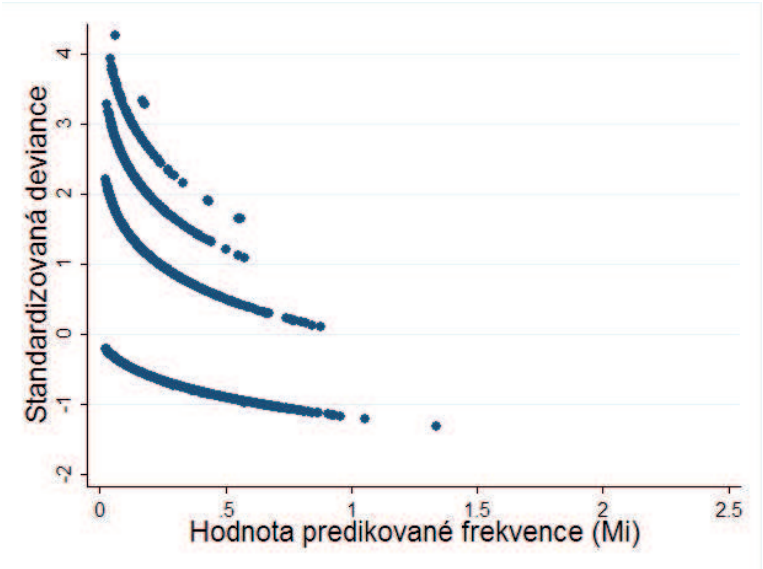
věrohodnosti nenabývá vysokých hodnot, jistá významnost kategorizace je zaznamenána. Pro lepší srovnání jsou zobrazeny grafy standardizovaných reziduí deviance modelu M1 na Obr. 4.21 a modelu Mi na Obr. 4.22.

Obr. 4.21 Srovnání predikované deviance modelu M1



Zdroj: vlastní zpracování v STATA 11.0

Obr. 4.22 Srovnání predikované deviance modelu Mi



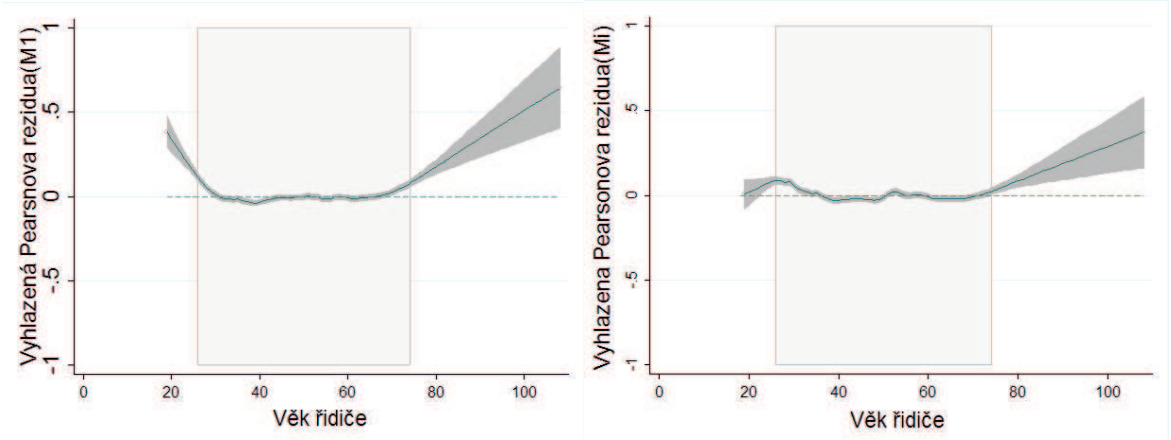
Zdroj: vlastní zpracování v STATA 11.0

Na Obr. 4.21 a Obr. 4.22 je zřetelné, že standardizovaná rezidua deviance jsou u modelu Mi významně menší. A požadavkem je, aby rezidua byla jak nejmenší. Rovněž

hodnota predikované frekvence je nižší v modelu s kategorickými veličinami. Proto model Mi s kategorickými veličinami je vhodnější pro odhad než model M1.

Dále je možnost porovnat vhodnost modelů pro modelování četnosti pomocí grafů vyhlazených Pearsnových reziduí. Nejdříve jsou Pearsnova rezidua predikována a následně vyhlazená vybranými vysvětlujícími proměnnými. Rezidua jsou vyhlazována pouze spojitými veličinami. Hodnota vyhlazených reziduí by měla být jak nejblíže horizontální ose v bodě nula. V níže uvedených grafech je vyznačena šedá plocha, kterou je určena oblast, kde by se mělo nacházet 95% pozorování. Šedou rozptýlenou oblastí kolem vývoje reziduí jsou představena všechna rezidua v 95 % intervalu spolehlivosti. Vyhlazená rezidua proměnnou věk řidiče modelu M1 a Mi jsou zobrazena na Obr. 4.23.

Obr. 4.23 Srovnání vyhlazených Pearsnových reziduí modelů M1 a Mi proměnnou věk řidiče

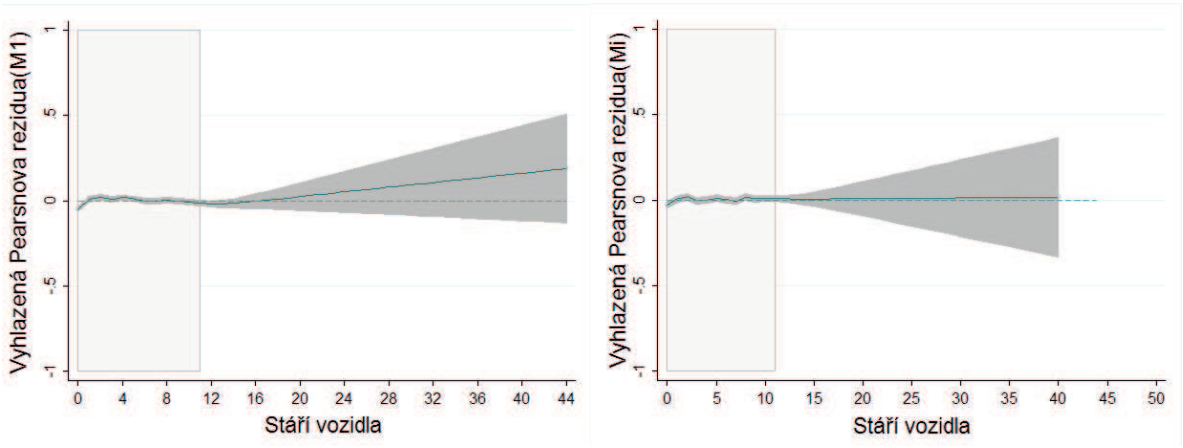


Zdroj: vlastní zpracování v STATA 11.0

Rezidua vyhlazená proměnnou věk řidiče se v obou případech vyvíjejí různě. V modelu M1 hodnoty reziduí u extrémních pozorování (mimo šedý obdélník) jsou mnohem více vzdálené od horizontální přímky v bodě nula oproti modelu Mi. Dalo by se říci, že rezidua vyhlazená proměnnou věk řidiče ani u jednoho modelu nejsou ideální. Ale u modelu s kategorickými veličinami extrémní hodnoty nejsou tak vzdálené od nuly.

Na Obr. 4.24 jsou srovnána rezidua vyhlazená proměnnou stáří vozidla.

Obr. 4.24 Srovnání vyhlazených reziduí modelů M1 a Mi proměnnou stáří vozidla

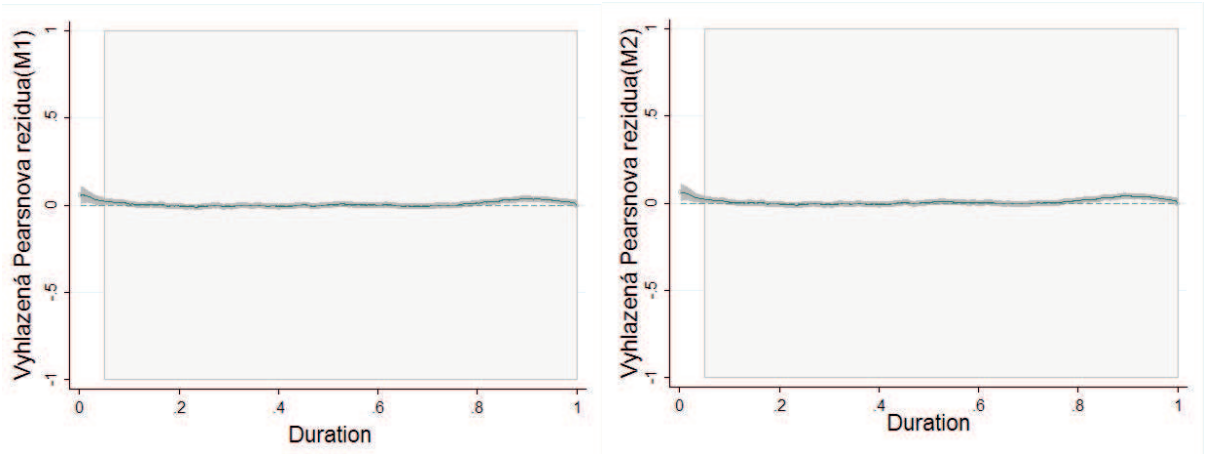


Zdroj: vlastní zpracování v STATA 11.0

V tomto případě již na první pohled je patrné, že vyhlazená rezidua modelu Mi včetně 95 % intervalu spolehlivosti jsou velmi blízká nule, respektive očekávané hodnoty reziduí odpovídající lineární funkci jsou významně blízké nule. Taktéž rezidua extrémních pozorování se významně neliší od nuly. Model Mi je mnohem vhodnější pro modelování četnosti pojistných událostí.

V následujícím Obr. 4.25 jsou zobrazena vyhlazená Pearsnova rezidua modelu M1 a Mi.

Obr. 4.25 Srovnání vyhlazených reziduí modelů M1 a Mi proměnnou duration



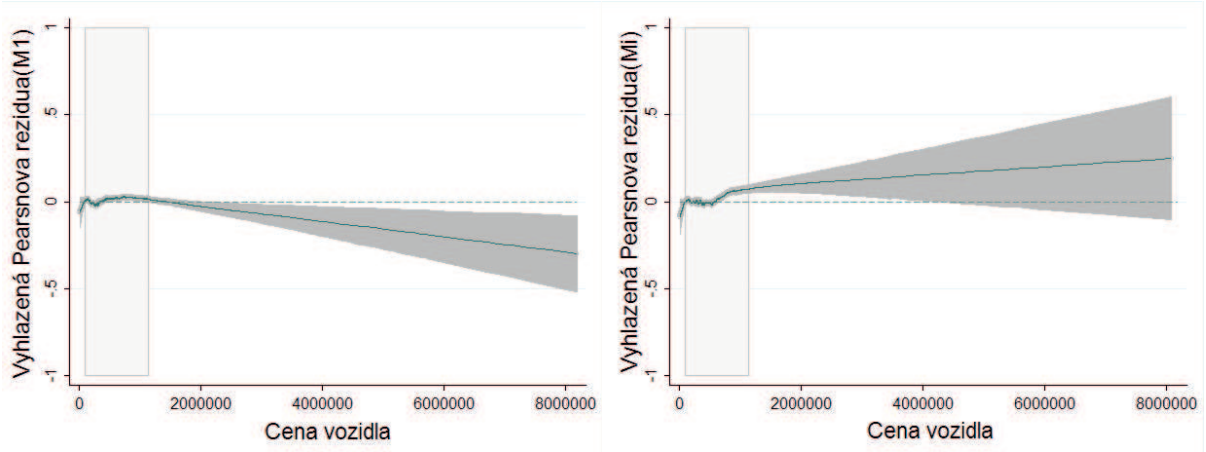
Zdroj: vlastní zpracování v STATA 11.0

Na základě vyhlazených Pearsnových reziduí proměnnou duration (časová expozice) nelze určit, který model je vhodnější. Hodnoty vyhlazených reziduí v obou případech jsou podobné ne-li shodné. Ale je možno konstatovat, že rezidui vyhlazenými proměnnou duration

je určena vhodnost odhadovaného negativně binomického modelu pro modelování četnosti škod.

Poslední srovnání reziduí je uvedeno na Obr. 4.26.

Obr. 4.26 Srovnání vyhlazených reziduí modelů M1 a Mi proměnnou cena vozidla



Zdroj: vlastní zpracování v STATA 11.0

Na Obr. 4.26 jsou zobrazena predikovaná Pearsnova rezidua model Mi s kategorizovanými veličinami a vyhlazená jsou proměnnou cena vozidla. V cenové relaci do 1 mil. Kč se nachází 95% pozorování. Vyhlazování touto proměnnou není vhodné, jelikož cenová relace je široká a nachází se tam rovněž vozidlo za víc jak 8 mil. Kč, čímž jsou výsledky zkreslené.

Hodnoty predikovaných reziduí u extrémních hodnot v modelu s kategorickými veličinami, mají mnohem nižší hodnoty než v odhadovaném modelu M1. Vyhlazená rezidua u modelu M1 a Mi se o moc neliší. Ale i přesto po testování poměrem věrohodnosti a grafickém srovnání je pro modelování četnosti pojistných událostí vhodnější pracovat s kategorizovanými veličinami. Důvodem vhodnosti použití pro modelování četnosti modelu Mi je to, že kategorickými veličinami je lépe vyjádřena nelineárnost modelu. Kategorizované proměnné jsou inherentně diskrétní proměnné a jsou používány v nelineárních pravděpodobnostních modelech. Při odhadování koeficientů modelu pomocí spojitých veličin je předpokládáno, že model je lineární. Ale ve výše odhadovaném modelu nejsou lineární funkce a kategorizace (transformace) spojitých proměnných lépe odpovídá nelinearitě. Odchylkami při modelování pomocí kategorických veličin je model lépe kopírován.

5 Závěr

Pojišťovnictví je jednou ze základních oblastí národního hospodářství. Pojištění je možné chápat jako přenesení rizika na jiný subjekt, pojišťovnu. Rizikem je myšlen vznik nahodilé událostí, která může mít negativní dopad na subjekt. Ten chce minimalizovat možné vzniklé škody, proto riziko přenáší na pojistitele. Základem pojistných teorií je rozdělení pravděpodobnosti a regresní modely.

Cílem diplomové práce bylo modelování četnosti pojistných událostí havarijního pojištění. Model byl odhadován metodou maximální věrohodnosti. Data, která byla k tomu použita, byla souborem dat týkajících se četnosti nahodilých událostí v havarijním pojištění za období od roku 2005 do roku 2010. K aplikaci metod odhadu na reálných datech byl použit statistický software STATA 11.0.

Ze základního souboru dat, po provedení jednofaktorové analýzy a testování multikolinearity mezi veličinami, byly vybrány proměnné do odhadovaného modelu. Zejména u kategorických veličin vyšlo, že na hladině spolehlivosti 95% některé kategorie vysvětlujících proměnných byly nevýznamné. Pro testování významnosti jako celku byl použit Waldův test a kategorické veličiny jako celek statisticky významné byly. A proto bylo rozhodnuto o jejich zařazení do odhadovaného modelu. Odhadovány byly koeficienty jak spojitých, tak i kategorických veličin. Model M1 byl odhadován metodou maximální věrohodnosti na hladině významnosti 5% a bylo předpokládáno, že náhodné veličiny pocházejí z negativně binomického rozdělení. Následně byl model zredukován o dvě proměnné a tento nový model M2 byl znova odhadován. Po porovnání těchto dvou odhadů vyšel významnější „nezredukovaný“ model M1. Kdyby z odhadu byly vynechány dvě proměnné, odhadovaný model pro modelování četnosti škod by byl nepřesný.

Další úkolem diplomové práce bylo zjištění, jaký vliv mají kategorické veličiny na modelování četnosti škod u havarijního pojištění. Všechny spojité veličiny z „nezredukovaného“ modelu M1 byly jednotlivě kategorizovány a následně podrobeny jednofaktorové analýze pro zjištění statistické významnosti odhadovaných parametrů. Všechny odhadované koeficienty vyšly na 5 % hladině významnosti statisticky významné, a proto byly zařazeny do odhadu kategorizovaného modelu Mi. Koeficienty modelu byly odhadovány metodou maximální věrohodnosti. Po odhadech byly modely M1 a Mi porovnány testem poměrem věrohodnosti. Jako mírně vhodnější vyšel model Mi s kategorickými vysvětlujícími proměnnými. U obou modelů byla rovněž porovnána

predikovaná rezidua a to jak rezidua deviance, tak i vyhlazená Pearsnova rezidua. V obou případech vyšla rezidua menší u odhadovaného modelu s kategorizovanými veličinami, to znamená, že model Mi je vhodnější pro modelování četnosti pojistných událostí havarijního pojištění.

Po odhadování modelů a následném grafickém hodnocení jsme došli k závěru, že negativně- binomická regrese je vhodná pro modelování četnosti pojistných škod u havarijního pojištění. I přes pár nevýhod, které se sebou nese kategorizování spojitých veličin, pro modelování četnosti nehod je lepší používat model s kategorickými veličinami. Důvodem jsou nelineární vztahy mezi proměnnými. Kategorickými veličinami jsou lépe znázorňovány odchylky od linearity v odhadovaném modelu. A existuje-li nelinearita v modelu, je lepší pro odhad použít kategorické veličiny.

Seznam použité literatury

1. Odborná literatura

CIPRA, Tomáš. *Finanční a pojistné vzorce*. Praha: Grada Publishing, 2006a. 376 s. ISBN 80-247-1633-X.

CIPRA, Tomáš. *Pojistná matematika- teorie a praxe*. 2. vyd. Praha: EKOPRESS, 2006b. 411 s. ISBN 80-86929-11-6.

HARDMAN, Joseph W. and J. M. HILBE. *Generalized Linear Models and Extensions*. 2nd ed. College station, Tex.: StataPress, 2007. ISBN 978-1-59718-014-6.

HILBE, Joseph M. *Negative binomial regression.* 2nd ed. New York: Cambridge University Press, 2011. ISBN 978-0521-19815-8.

JONG, Piet de and Gillian Z. HELLER. *Generalized Linear Models for Insurance Data*. Cambridge: Cambridge University Press, 2008. 196 s. ISBN 978-0-521-87914-9.
OHLSSON, Esbjörn and Björn JOHANSSON. *Non-Life Insurance Pricing with Generalized Linear Models*. Berlin: Springer, 2010. 174 s. ISBN 978-3-642-10790-0.

ROYSTON, Patrick and W. SAUERBREI. *Multivariable model-building: a pragmatic approach to regression analysis based on fractional polynomials for modelling continuous variables*. Hoboken, NJ: John Wiley, c2008. ISBN 978-0-470-02842-1.

2. Odborné články

BRANDA, Martin. Zobecněné lineární modely v pojišťovnictví. *Zpracováno v rámci projektu Fondu pro podporu vzdělávání v pojišťovnictví.* 2013. 34 s. Praha.

VALECKÝ, Jiří. Modelling claim frequency in vehicle insurance. *Acta Universitatis Agriculturae et Silviculturae Mendelianae Brunensis.* 2015. 10 s. ISSN 1211-8516.

VALECKÝ, Jiří. Modelling of individual insured accident risk for given motor- hull insurance portfolio. *The 6th International Days of Statistics and Economics, Prague, September 13-15*. 2012a. s. 1143-1151. ISBN 978-80-86175-86-7.

VALECKÝ, Jiří. Fractional polynomials analysis of relation between insured accident and selected risk factors. *Proceedings of 30th International Conference Mathematical Methods in Economics*. 2012b. s. 956-961. ISBN 978-80-7248-779-0.

Seznam zkratek

AIC	Akaikeho informační kritérium
BIC	Bayesovo informační kritérium
EIM	očekávaná informační matice
GLM	zobecněný lienární model
IRLS	metoda iterativně vážených nejmenších čtverců
LR	test poměrem věrohodnosti
mil.	miliony
MLE	odhad maximální věrohodnosti
MM	metoda momentů
NB	negativně- binomické rozdělené
NB2	typ negativně- binomického rozdělení s kvadratickou funkcí rozptylu
OIM	pozorovaná informační matice
tis.	tisíce

Prohlášení o využití výsledků diplomové práce

Prohlašuji, že

- jsem byla seznámena s tím, že na mou diplomovou práci se plně vztahuje zákon č. 121/2000 Sb. – autorský zákon, zejména § 35 – užití díla v rámci občanských a náboženských obřadů, v rámci školních představení a užití díla školního a § 60 – školní dílo;
- beru na vědomí, že Vysoká škola báňská – Technická univerzita Ostrava (dále jen VŠB-TUO) má právo nevýdělečně, ke své vnitřní potřebě, diplomovou práci užít (§ 35 odst. 3);
- souhlasím s tím, že diplomová práce bude v elektronické podobě archivována v Ústřední knihovně VŠB-TUO a jeden výtisk bude uložen u vedoucího diplomové práce. Souhlasím s tím, že bibliografické údaje o diplomové práci budou zveřejněny v informačním systému VŠB-TUO;
- bylo sjednáno, že s VŠB-TUO, v případě zájmu z její strany, uzavřu licenční smlouvu s oprávněním užít dílo v rozsahu § 12 odst. 4 autorského zákona;
- bylo sjednáno, že užít své dílo, diplomovou práci, nebo poskytnout licenci k jejímu využití mohu jen se souhlasem VŠB-TUO, která je oprávněna v takovém případě ode mne požadovat přiměřený příspěvek na úhradu nákladů, které byly VŠB-TUO na vytvoření díla vynaloženy (až do jejich skutečné výše).

V Ostravě dne 22. dubna 2016

Ester Lysková

Ester Lysková

Seznam příloh